Dissecting the genetic architecture of Multiple Sclerosis

By

Rui Lin

B.A (Preventive Medicine; China)
M.D (Pathogen Biology; China)

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

UNIVERSITY OF TASMANIA

University of Tasmania
April 2014

Declaration of originality

This thesis contains no material which has been accepted for any degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my acknowledge and belief, no material previously published or written by another person, except where due acknowledgment is made in the text of the thesis.

Signature

18 April 2014

Rui Lin | Date

Statement of authority to access

This thesis may be available for loan. Copying of any part of this thesis is prohibited for two years from the date this statement was signed; after that time, limited copying is permitted in accordance with the *Australian Copyright Act of 1968.*

Signature

18 April 2014

Rui Lin — Date

Statement regarding published work contained in thesis

The publishers of the papers comprising Chapters 1 to Chapter 6 hold the copyright for that content, and access to the material should be sought from the respective journals. The remaining non-published content of the thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Signature

18 April 2014

Rui Lin Date

Statement of Co-authorship

This thesis includes papers for which Rui Lin (RL) is not sole author. RL took the lead in this research, developing and implementing the analyses, writing manuscript included herein under the supervision of Bruce V Taylor (BVT), Jac Charlesworth (JC), Ingrid van der Mei (IvM) and Steve Simpson, Jr. (SSJ). In this process, however, she was assisted by co-authors to varying extent. Following then, the contribution of each co-author is detailed for each respective project.

The paper reported in Chapter 1:

Rui Lin, Jac Charlesworth, Ingrid van der Mei, Bruce V Taylor. "The genetics of multiple sclerosis". *Practical Neurology.* 2012 Oct;12(5):279-88. (Invited review)

RL undertook the literature review with direct assistance from all authors. All authors contributed equally to the development of the review.

The paper reported in Chapter 2:

Rui Lin, Jac Charlesworth, Jim Stankovich, Victoria M. Perreau, Matthew A. Brown, ANZgene Consortium, Bruce V Taylor. "Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis". *PLoS One*. 2013;8(3):e56379. doi: 10.1371/journal.pone.0056379. Epub 2013 Mar 5.

RL contributed to data analyses undertaken in this study and composed drafts of the manuscript and coordinated revision.

Jim Stankvich (JS) conceived and designed this study, and provided guidance and supervision for all statistical analyses undertaken in this study, and was involved in the critical revision of the manuscript.

Victoria M. Perreau (VMP) was involved in the system biology analysis and critical revision of the manuscript.

JC, Matthew A. Brown (MAB) and BVT contributed to the critical revision of the manuscript.

ANZgene Consortium provided access to the data relating to the MS associated SNPs and was involved in critical revision of the manuscript.

In chapter 3:

RL was involved in the development and implementation of statistical analyses undertaken, under supervision by JC.

BVT and JC conceived and designed this study. JC contributed re-analyses the raw whole genome sequencing data using the pedigree-based pipeline in Biomedical Research Institute, Texas, USA.

Wet-lab Whole genome sequencing (WGS) libraries were prepared and sequenced by Marina Donskoi and Sharon Song under the direction of Brooke Gardiner in Queensland University.

The paper reported in Chapter 4:

Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. "Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis". *Journal of Neurology, Neurosurgery & Psychiatry*. 2014 Apr;85(4):399-404. doi: 10.1136/jnnp-2013-305245. Epub 2013 Jul 18.

RL was involved in the development and implementation of statistical analyses undertaken, under supervision by SSJ, IvM, JC and BVT. RL composed drafts of the manuscript and coordinated revision.

BVT was involved in the development and acquisition of funding for the MS Longitudinal Study from which the data for this analysis was drawn, along with Anne-Louise Ponsonby (A-LP), Fotini Pittas (FP), Terence Dwyer (TD) and IvM. BVT was involved in the critical revision of the manuscript.

SSJ provided guidance and supervision for the epidemiological statistical analyses undertaken in this study, and was involved in the critical revision of the manuscript.

JC provided guidance and supervision for the genetic statistical analyses undertook in this study, and was involved in the critical revision of the manuscript.

A-LP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, FP, TD, and IvM. A-LP contributed to the critical revision of the manuscript.

FP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, TD, and IvM. FP was involved in the data collection for the MS Longitudinal Study along with BT and IvM. FP contributed to the critical revision of the manuscript.

TD was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and IvM. TD contributed to the critical revision of the manuscript.

IvM was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and TD. IvM was involved in the data collection for the MS Longitudinal Study along with BT and FP. IvM was involved in conception of the analyses used. IvM was involved in the critical revision of the manuscript.

The paper reported in Chapter 5:

Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. “Association between multiple sclerosis risk-associated SNPs and relapse and disability - a prospective cohort study.” *Multiple Sclerosis.* 2014 Mar;20(3):313-21. doi: 10.1177/1352458513496882. Epub 2013 Jul 25.

RL was involved in the development and implementation of statistical analyses undertaken, under supervision by SSJ, Ingrid A F van der Mei (IvM), JC and BVT. RL composed drafts of the manuscript and coordinated revision.

BVT was involved in the development and acquisition of funding for the MS Longitudinal Study from which the data for this analysis was drawn, along with Anne-Louise Ponsonby (A-LP), Fotini Pittas (FP), Terence Dwyer (TD) and IvM. BVT was involved in the critical revision of the manuscript.

SSJ provided guidance and supervision for the epidemiological statistical analyses undertaken in this study, and was involved in the critical revision of the manuscript.

JC provided guidance and supervision for the genetic statistical analyses undertaken in this study, and was involved in the critical revision of the manuscript.

A-LP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, FP, TD, and IvM. A-LP contributed to the critical revision of the manuscript.

FP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, TD, and IvM. FP was involved in the data collection for the MS Longitudinal Study along with BT and IvM. FP contributed to the critical revision of the manuscript.

TD was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and IvM. TD contributed to the critical revision of the manuscript.

IvM was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and TD. IvM was involved in the data collection for the MS Longitudinal Study along with BT and FP. IvM was involved in conception of the analyses used. IvM was involved in the critical revision of the manuscript.

The paper reported in Chapter 6:

Rui Lin, Bruce V Taylor, Jac Charlesworth, Ingrid van der Mei, Leigh Blizzard, Niall Stewart, Anne-Louise Ponsonby, Terry Dwyer, Fotini Pittas, Steve Simpson, Jr.. "Modulating effects of the gene *WT1* on the interferon-β-vitamin D association in MS" (submitted).

RL was involved in the development and implementation of statistical analyses undertaken, under supervision by SSJ and JC. RL composed drafts of the manuscript and coordinated revision.

BVT was involved in the development and acquisition of funding for the MS Longitudinal Study from which the data for this analysis was drawn, along with Anne-Louise Ponsonby (A-LP), Fotini Pittas (FP), Terence Dwyer (TD) and IvM. BVT was involved in the critical revision of the manuscript.

JC provided guidance and supervision for the genetic statistical analyses undertook in this study, and was involved in the critical revision of the manuscript.

IvM was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and TD. IvM was involved in the data collection for the MS Longitudinal Study along with BT and FP. IvM was involved in the critical revision of the manuscript.

Leigh Blizzard (LB) was involved in the development of double interaction analyses undertaken in this study, and was involved in the critical revision of the manuscript.

Niall Stewart (NS) was involved in the critical revision of the manuscript.

A-LP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, FP, TD, and IvM. A-LP contributed to the critical revision of the manuscript.

FP was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, TD, and IvM. FP was involved in the data collection for the MS Longitudinal Study along with BT and IvM. FP contributed to the critical revision of the manuscript.

TD was involved in the development and acquisition of funding for the MS Longitudinal Study, along with BT, A-LP, FP, and IvM. TD contributed to the critical revision of the manuscript.

SSJ conceived this study, and provided guidance and supervision for the epidemiological statistical analyses undertaken in this study, and was involved in the critical revision of the manuscript.

Signed _____

Professor Bruce Taylor

Supervisor

Menzies Research Institute Tasmania

University of Tasmania

Date _____

Signed _____

Professor Tom Marwick

Head of School

Menzies Research Institute Tasmania

University of Tasmania

Date _____

Statement of Ethical Conduct

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University.

Signature

18 April 2014

Rui Lin

Date

Acknowledgement

Sincere thanks my supervisors, Bruce, Jac, Ingrid, Steve and Jim, for their thorough and patient guidance and support throughout the duration of my PhD. They provided me the structure within which I could complete my projects, while giving me the intellectual freedom to develop my own ideas and projects. Meanwhile, as an international student, I am deeply grateful to my supervisors, for their constant help to improve my English. I am grateful to the MS group and the genetics group in Menzies, for their useful comments when I made the lab-meeting presentation, and for a pleasant and inspiring atmosphere in these two groups.

Outside my work world, I sincerely think my family's support, particularly my dear husband, Shunpei. He keeps giving me smiles no matter what had happened, even my father had a big surgery in China and my father-in-law passed away during my PhD study. Without his selfless support and dedication, I could not finish my PhD on time. My sincere thanks go also to Mr Yuan Zhou, for his help in printing my thesis after I came back China.

Lastly, I will offer my thanks to the participants of the MS Longitudinal Study, as well as the participants in whole genome sequencing project. So at the very least I hope that the work of my thesis in some way is of help to our participants, and extend to all persons.

Abstract

Multiple sclerosis (MS) is an inflammatory demyelinating disease of the central nervous system (CNS) triggered by environmental and genetic factors. There is considerable evidence for a significant genetic component to MS susceptibility, such that genome-wide association studies (GWASs) have successfully revealed about 100 common genetic variants associated with MS. However, at least 60% of the overall heritability of MS has not been delineated, so-called “missing heritability”. This thesis describes a series of experiments to further elucidate the “missing heritability” of MS.

Chapter 2 describes how to detect rare variants conferring susceptibility to MS using identity-by-decent (IBD) mapping, utilising a large GWAS dataset comprising 3,543 cases and 5,898 controls. We identified a genome-wide significant linkage signal on chromosome 19. This linkage region includes a cluster of zinc finger genes that were suggested may be involved in very early developmental regulation of the CNS via genome wide transcriptome analysis.

Chapter 3 describes how to detect rare or disease-causing variants to MS by employing whole genome sequencing on eight samples from one extended pedigree. We identified nine candidate variants by both whole genome sequencing analysis and Sanger resequencing. Of these, *RDBP* and *PKHD1* are the two most interesting genes, which have been demonstrated to be associated with MS and ALS previously. Further research on these two genes will be continued after this thesis is submitted.

Chapter 4 and Chapter 5 describes how to detect gene-environment interactions particularly gene-vitamin D interactions influencing MS clinical course via a well-validated MS prospective cohort study followed from 2002 to 2005. In Chapter 4 we find the relationship between 25(OH)D and hazard of relapse was significantly different for different alleles of two intronic SNPs (rs908742 in *PRKCZ* and rs3783785 in *PRKCH*) in the protein kinase C (PKC) family genes. Additionally, in Chapter 5 we find a number of known MS risk-associated SNPs were associated with relapse, or altered the 25(OH)D-relapse association or modified levels of 25(OH)D. Our findings indicate gene-vitamin D interactions may be an important mechanism on MS clinical course, and provide support for the role of vitamin D in MS relapse.

In this MS cohort, it was also found that *WT1* variants may play a role in altering the therapeutic effects of IFN-β in MS, such that two SNPs (rs10767935 and rs5030244) in *WT1* gene significantly modified the previously demonstrated association between IFN-β and 25(OH)D in the patients with MS.

This thesis presents a range of studies which add significantly to the literature on MS genetics as well as MS epidemiology. This work will be useful in the scientific community; both for hypothesis generation and providing strong evidence in support of existing hypothesis, and hopefully be of benefit to people with MS.

Table of contents

List of tables

List of figures

Papers directly arising from the work described in this thesis

Paper published

Chapter 1: Rui Lin, Jac Charlesworth, Ingrid van der Mei, Bruce V Taylor. The genetics of multiple sclerosis. *Practical Neurology 2012;12:279-288* (Appendix 1A) (invited review).

Chapter2: Rui Lin, Jac Charlesworth, Jim Stankovich, Victoria M. Perreau, Matthew A. Brown, ANZgene Consortium, Bruce V Taylor. Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis. *PLoS One*. 2013;8(3):e56379 (Appendix 2A).

Chapter 4: Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*. 2014; 85(4):399-404 (Appendix 4A).

Chapter 5: Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. Association between multiple sclerosis risk-associated SNPs and relapse and disability - a prospective cohort study. *Multiple Sclerosis.* 2014; 20(3):313-21. (Appendix 5A).

Conference presentations arising from work in this thesis

Oral presentation:

2011 Identity-by-descent mapping to identify rare variants conffering susceptibility of multiple sclerosis. The 18th GeneMapper Conference. 4-6 April 2011, Hobart.

Poster presentations:

2011 Searching for rare variants conferring susceptibility to multiple sclerosis. The MSRA Progress in MS Research Conference. 26-28 October, 2011, Melbourne.

2012 Graduate Research-Sharing Excellence in Research Conference, 6-7 September 2012, Hobart, University of Tasmania.

Awards received from the work described in this thesis

Travel bursary to attend the BioInfoSummary course. December 2010, Melbourne.

List of abbreviations

MS	Multiple sclerosis
CNS	Center nervous system
EDSS	Expanded Disability Status Scale
MSSS	Multiple Sclerosis Severity Score
NRS	Scripps Neurologic Rating Scale
RRMS	The relapsing remitting MS
PPMS	The primary progressive MS
UVR	Ultraviolet radiation
25(OH)D	25-hydroxyvitamin D
$1,25(OH)_2D_3$	1,25-dihydroxyvitamin D
VDR	Vitamin D receptor
RXR	retinoid X receptor
EAE	Experimental autoimmune encephalomyelitis
EBV	Epstein-Barr virus
EBNA	EB nuclear antigen
MHC	The major histocompatibility complex
HLA	The human leukocyte antigen
GWAS	Genome-wide association study
SNP	Single-nucleotide polymorphism
nsSNPs	nonsynonymous SNPs
MAF	Minor allele frequency
LD	Linkage disequilibrium
IMSGC	The International Multiple Sclerosis Genetics Consortium
WTCCC2	Wellcome Trust Case Control Consortium 2

ANZgene	The Australia and New Zealand Multiple Sclerosis Genetics Consortium
IBD	Identity-by-descent
PBLA	Population-based linkage analysis
PCA	Principal components analysis
IPA	Ingenuity Pathway Analysis
BP	Base pair
NGS	Next-generation sequencing
SNV	Single nucleotide variant
GERP	Genomic Evolutionary Rate Profiling
SIFT	Sorting Tolerant From Intolerant
Polyphen2	Polymorphism Phenotying v2
IGV	Integrative Genomics Viewer
ALS	Amyotrophic lateral sclerosis
HR	Hazard ratio
OR	Odds ratio
CI	Confidence interval
BMI	Body mass index
PKC	Protein kinase C
IFN- β	Interferon-beta
SNPSpD	Single Nucleotide Polymorphism Spectral Decomposition
MeffLi	The effective number of independent marker loci

Other publication

O'Gorman C, **Lin R,** Stankovich J, Broadley SA. Modelling genetic susceptibility to multiple sclerosis with family data. *Neuroepidemiology*. 2013; 40(1):1-12 (Appendix 7A).

Chapter 1. Background: multiple sclerosis and its genetics

1.1 Preface

This chapter gives an overview of the disease multiple sclerosis (MS): its clinical features, pathogenesis, immunomodulatory treatment, and risk factors, specifically the genetic contribution to MS susceptibility. It provides information relevant for the understanding of the thesis. The part, "The genetics of multiple sclerosis" has been published in *Practical Neurology 2012;12:279-288* (Appendix 1A). The grey boxes within this part are added and were not part of the original publication.

1.2 Multiple sclerosis

Multiple sclerosis (MS) is an inflammatory demyelinating disease of the central nervous system (CNS) resulting in damage to myelin sheaths and CNS axons, although the underlying pathophysiology and precise aetiology remains unknown. MS affects 2.5 million individuals worldwide of which most are young adults of northern European ancestry [1]. The prevalence in southern Australia and New Zealand exceeds 1 per 1000 and is more common in females than in males, with a sex ratio exceeding 3:1. The median age of MS onset is in the fourth decade of life and there is no known cure, with the disorder leading to progressive disability in the majority of patients and a large economic burden in the community. Improved therapy and prevention depend on a better understanding of the causes and mechanisms involved in the onset and progression of MS.

1.3 Clinical features of MS

People with MS can suffer from a broad spectrum of signs and symptoms. The most common initial symptoms include muscle weakness and spasms, sensory disturbances, visual problems and ataxia (difficulties with coordination and balance). Additional features that develop with the disease course are bladder and bowel dysfunction, fatigue, dysarthria (problems in speech), dysphagia (problems in swallowing) and pain [2]. It has been recognised that cognitive impairment, concentration, depression and unstable mood are also common during the course of the disease [2]. Several measures have been implemented to clinically assess disability progression and symptom severity including the Kurtzke Expanded Disability Status Scale (EDSS), Multiple Sclerosis Severity Score (MSSS) and Scripps Neurologic Rating Scale (NRS). The EDSS is the gold standard in assessing physical disability in MS, which utilises a ten-point disease severity scale from 0 (normal) to 10 (death due to MS). Scores of 1-4 usually represent the symptoms and physical severity in range from no disability & minimal signs in one Function System to moderate disability. Scores of 4-5.5 generally reflect limiting walking 500 meters to 100 meters without aid or rest. Scores over 6 range from requiring assistance in walking to being totally helpless in bed [3]. MSSS is a useful measure of MS severity. It can be viewed as a cross-sectional measure of disability progression, as it uses the EDSS and takes disease duration into account. It allows a rapid assessment of relative ranks of individuals from lowest EDSS to highest EDSS for a given disease duration, and expresses this as a decile rank between 0 (least affected) and 10 (most severely affected) [4]. NRS score is based on assessment of each component of the neurologic examination including 22 parameters of motor, sensory and cognitive function, and accurately reflects

overall neurologic function, ranging from 100 (Normal) to 10 (Maximal impairment) [5].

MS is classified into three subtypes based on the United States National Multiple Sclerosis Society standard: relapsing remitting, secondary progressive and primary progressive (**Figure 1.1**) [6]. The relapsing remitting form (RRMS) is characterised by unpredictable relapses followed by remission with return to or nearly to pre-existing levels of disability, followed by variable periods of disease quiescence. This subtype is the predominant form of MS, and accounts for >80% of cases in most case series [2]. The average age of onset of RRMS is in the 3th decade of life [2]. Most cases of RRMS will convert to secondary progressive MS (SPMS) over time, >50% doing so by 15 years of disease duration, this course defined by a gradual worsening of function rather than new relapses. The primary progressive MS (PPMS) form is a condition accounting for 10–15% of individuals, and is characterised by progression of disability from onset without any relapses or remission [7]. The age of onset for the primary progressive subtype is later than for the relapsing-remitting, generally around 40 years of age [8]. In some occasions a fourth type of MS clinical course is described, of relapsing progressive MS, where progression occurs from onset with superimposed relapses.

Figure 1.1. Types of MS.

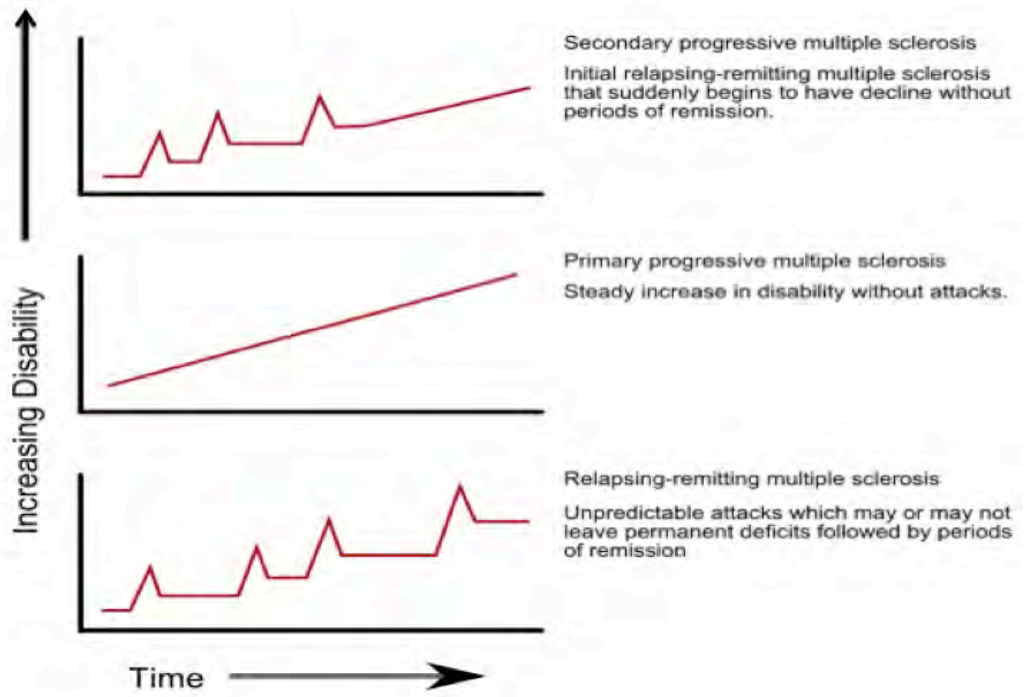


Figure modified from Lublin F.D., et al. [9]

1.4 Pathogenesis of MS

MS is a chronic inflammatory disorder of the brain and spinal cord, leading to myelin and axonal damage caused by the innate and adaptive inflammatory responses. Myelin is synthesised by mature oligodendrocytes, each of which contact axons in white matter tracts of the CNS. The disease process starts with increased migration of autoreactive lymphocytes: T-cells and B-cells cross into the blood-brain barrier, and their co-stimulation with other immune cells such as interleukin-17 producing (Th17 cells) [10] and soluble factors, including cytokines and antibodies, which in turn leads to demyelination, oligodendrocyte death and subsequent axonal loss. Notably, Th17 has

been considered the driver of inflammation under Th23 control rather than Th1, due to sufficient Th17 cells crossing the blood-brain barrier to kill human neurons and promote CNS inflammation through CD4+ lymphocyte recruitment [11]. MS lesions commonly involve white matter area whose function is to carry signals between grey matter area. Oligodendrocytes are responsible for creating and maintaining the myelin sheath to allow saltatory conduction along axons. With disease onset, myelin can be thinned or completely lost, such that an axon can display partial or complete conduction block [12]. Additionally, as T lymphocytes, B lymphocytes, plasma cells, and macrophages accumulate, pro-inflammatory cytokines amplify the immune response through recruitment of naive microglia. On activation, these cells increase surface binding for myelin basic protein, show greater contact with living oligodendrocytes and produce TNF-α, which is capable of killing oligodendrocyte and the activated microglia phagocytose myelin breakdown products [13].

Demyelinating lesions, also called plaques, cluster around the lateral ventricles and corpus callosum, in the cortex and subcortical white matter, the optic nerves and brainstem, and throughout the spinal cord. They are characterized by perivascular CD8+cell infiltrates [2]. Acute demyelinating lesions also demonstrate extensive axonal injury, potentially mediated by CD8+ cytotoxic T-cell and microglial infiltration [14]. In the secondary progressive stage, demyelination coexists with diffuse axonal and neuronal degeneration. Cortical demyelination and injury of the normal-appearing white matter are reflected by diffuse axonal injury with profound microglia activation, which occur on the background of a global inflammatory response in the whole brain and meninges [15]. Pathological changes of primary progressive MS are characterised by reduced plaque load, less evidence for inflammation, and the absence of lymphoid follicles. Completely

remyelinated lesions and partially remyelinated lesions are also called shadow plaques. Remyelination is most active during the acute inflammatory process, coinciding with phagocytic removal of myelin debris, but also occurs in the progressive phase. In 20% of people with MS, plaques are eventually remyelinated [16].

1.5 Immunomodulatory treatments for MS

To date, no effective curative agent has been developed since the pathogenesis of MS is not adequatly understood. As MS is believed to be the result of a misdirected immune response [17], all treatments aimed at treating MS are immunomodulatory in nature, either those derived from immunomodulatory cytokines (interferon-β) and immune components (antibody therapies), and compounds which modulate the immune system (glatiramer acetate, corticosteroids, cladribine, azathioprine, mitoxantrone). In practice, treatment varies with the stage reached in the course of the disease, but mainly aim to reduce relapse rate and the accumulation of disability. Amongst the available therapeutic agents, reducing the relapse rate is best shown by use of the interferon-β (IFN-β), which has been in widespread use as first-line disease modifying agents in RRMS since the early 1990s. These include interferon β-1a (Avonex, Biogen, and Rebif, Ares-Serono) and interferon β-1b (Betaferon and Betaseron, Schering). The type-1 interferons generally are considered to have an immunomodulatory mechanism of action: IFN-β inhibits the proliferation of T lymphocytes and reduces their production of IFN-γ [18] which regulates the entry of T-cells into the CNS [19], and it alters the profile of cytokine production toward that of the anti-inflammatory Th2 response and reduces T-cell migration by inhibiting the activity of T-cell matrix metalloproteinases [19], as well as reducing antigen presentation [20].

Interferon-βs have been studied in separate placebo-controlled trials with RRMS. It has been demonstrated that interferon β-1a treatment can produce a significant delay in time to sustain EDSS progression, fewer exacerbations and fewer new lesions over 2 years [21] or 4 years [22]. A lower relapse rate and a delay to first relapse by 3 and 5 months has been observed, as well as an increased proportion of relapse-free patients [23]. Interferon β-1b treatment showed median 80% reduction in Magnetic Resonance Imaging (MRI) activity, and a significant reduction in relapse frequency and severity [24]. There was a one-third reduction in exacerbation rate in the 8- million international units (MIU) treatment arm in those patients with interferon β-1b use, compared with placebo, over 5 years [25].

Besides interferon-βs, glatiramer acetate is another first-line agent for use in the treatment of RRMS. The mechanisms of glatiramer acetate remain incompletely understood. It is now thought that glatiramer acetate modifies the immune response in many ways, including inducing tolerance of myelin-reactive T-cells, increasing regulatory T-cells, inhibiting the activation of myelin basic protein (MBP)-specific T-cells, producing IFN-gamma and inducing highly cross-reactive Th2-type T-cells [26,27]. From 1996, glatiramer acetate has been approved as first line treatment for RRMS in United States and currently is approved in Australia and New Zealand as well. It has been demonstrated that glatiramer acetate can increase the proportion of patients remaining relapse-free [28], reduce relapse rate [29], MRI activity and brain atrophy [30,31], and delay the progression from clinically isolated syndrome (CIS) to clinically definite MS [32]. However, a side effect was observed in the use of glatiramer acetate, and no effect in progressive MS [33].

At present, least contentious is the use of the IFN-β and glatiramer acetate in relapsing-

remitting disease. Studies showed a modest effect on the accumulation of disability with interferon-β treatment over 5 years [34] or 3 years [35]. These therapies have yet had no useful treatment effect on the secondary progressive phase of the disease [36-38]. Mitoxantrone, an anthracenedione antineoplastic drug that intercalates with DNA and inhibits both DNA and RNA synthesis, is considered probably more efficacious than the interferons or glatiramer acetate, but its use is confined to patients with aggressive relapsing-remitting MS due to its significant side effect profile with cardiac toxicity and increased rates of leukaemia following its use [39]. In particular, it slows the accumulation of disability in very active MS cases, but has less or no effect on non-relapsing progressive disease [40,41]. No agent has been shown to exert effects on primary progressive MS [42].

Other newly developed immunosuppressive therapies include natalizumatb, fignolimod and teriflunomide have been reported with superior efficacy in MS in recent years. Natalizumab is a humanised monoclonal antibody to α4 integrin (VLA-4), which is expressed predominantly on lymphocytes, monocytes, eosinophils, and basophils and is critical in mediating Th-1 cell migration in EAE [43-45]. Fingolimod is a derivative of myriocin, and has immunosuppressive qualities through modulation of the sphingosine 1 phosphate (S1P) system [46]. Teriflunomide selectively and reversibly inhibits the mitochondrial enzyme dihydroorotate dehydrogenase, which is required for *de novo* pyrimidine synthesis in proliferating lymphocytes [47]. A 68% reduction in relapses and 42% reduction in the risk of sustained progression of disability has been observed with use of natalizumab, compared with placebo in a pivotal clinical trial [48]. In addition, an 83% reduction in the effects on MRI T2 and gadolinium-enhancing lesions was observed over two years [48]. As for fingolimod, two large phase III trials have been reported. Both

trials showed the superior efficacy of oral fingolimod in most MRI outcomes, and with statistically significant reduction of relapse rate for both doses compared with either IFN-β [49] or placebo [50]. Similarly, in a phase III clinical trial, teriflunomide has been demonstrated to reduce annualised MS relapse rate and the risk of sustained disability progression by 31.5-36.3% and approximately 25-30% versus placebo, respectively, besides the positive effects on MRI outcomes [51]. Side effects in use of these new therapies were reported, including hypersensitivity reactions and the risk of progressive multifocal leukoencephalopathy (PML) for natalizumab, cardiac toxicity for fingolimod, and nausea, diarrhea, hair thinning and predominantly asymptomatic elevation of alanine aminotransferase levels for teriflunomide.

Recently, a number of other agents have emerged and are likely to become part of the MS treatment armament within the next few years [52], including BG-12 (dimethyl fumarate), alemtuzumab, daclizumab, ocrelizumab and laquinimod. In a phase IIb trial and two phase III trials, BG-12 has been shown to produce a reduction in annual relapse rate (approximate 50%), and in gadolinium-enhancing lesions (90%) and in new T1 hypointense lesions (60-70%) [53-55]. The mechanism of BG-12 is thought to act through activation of the Nrf-2 pathway, an important protective pathway against oxidative stress [56]. Alemtuzumab, a monoclonal antibody to CD-52, has been shown to be highly effective in MS trials compared to active therapy (interferon- β-1a) [57-59]. All three clinical trials demonstrated significant reductions in relapse rate (49-74%), and two showed a significant reduction in sustained accumulation of disability (42-65%) [58,59]. As for daclizumab, a monoclonal antibody to CD-25, several small trials (patient numbers 9 to 19) have shown it is effective in reducing contrast enhancing lesions and relapse rates as an add-on therapy [60] or as monotherapy [61] or in overlapping algorithms depending on response to therapy [62,63]. A larger phase IIb study with 230 patients also

demonstrated that contrast enhancing lesions were reduced by 72% in the higher dose [64]. Ocrelizumab, a humanised anti-CD20 monoclonal antibody, has demonstrated a significant reduction in new gadolinium-enhancing lesions by 89% and 96% respectively at doses of 600mg and 2000mg in a phase II trial, and the relapse rate was reduced by between 50-70% at 24 weeks [65]. Although positive outcomes have been demonstrated by the use of these emerging therapies, further assessments in large, long-term trials are warranted.

Treatment guidelines for MS in adults in Australia and New Zealand recommend that patients with active relapsing-remitting disease (2 relapses in 2 years) should be offered interferon-β, glatiramer acetate, teriflunomide, fingolimod or natalizumab, and clinical or MRI relapses should prompt consideration of escalation in therapy to fingolimod or natalizumab. When there are poor prognostic indicators in RRMS from the outset, fingolimod or natalizumab are recommended as first line medication. However, there is no effective therapy in primary progressive MS or established secondary progressive disease currently. The issue of whether therapy should be discontinued in those converting to progressive disease on a background of relapses is complex. Where there is clinical or MRI evidence of active inflammation, any continuous therapy seems to have continued effectiveness. While once a state of severe disability has been achieved there is probably very little to be gained from continuing therapy, however.

1.6 Environmental factors and MS onset

There is considerable evidence that MS is driven by both environmental and genetic factors, with the known environmental factors including lower ultraviolet radiation (UVR)

exposure and/or lower vitamin D levels, Epstein Barr virus, and smoking now well established.

1.6.1 UVR exposure and vitamin D

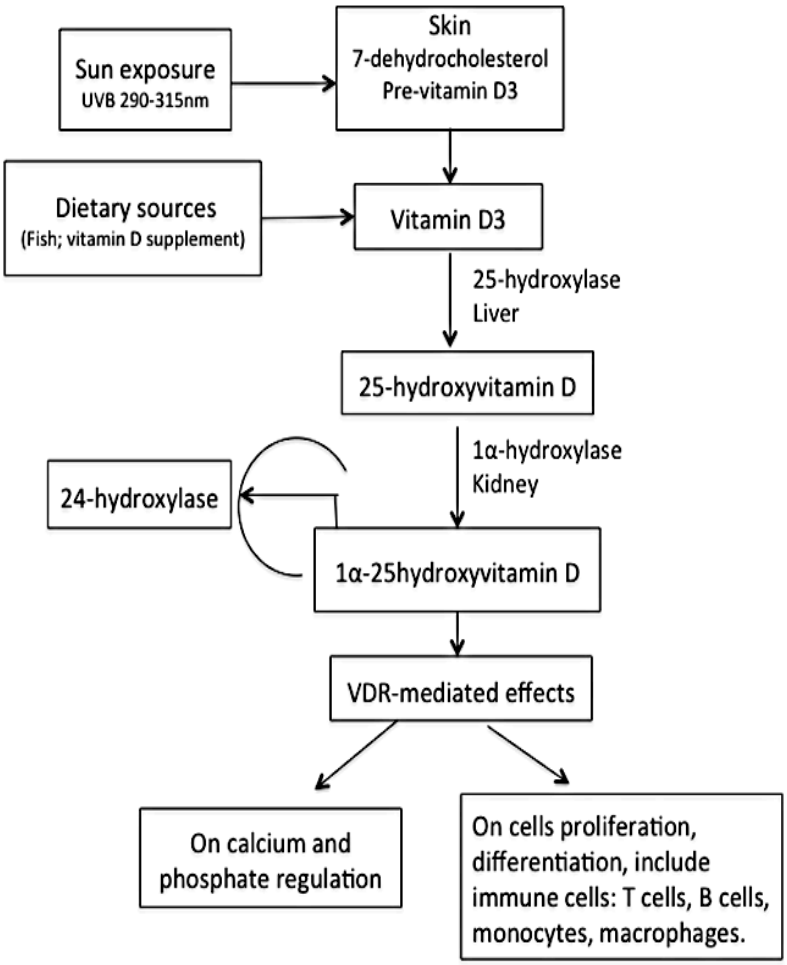
Vitamin D deficiency has long been considered as a risk factor for MS. Recently, with the immunomodulatory effects of vitamin D [66] along with increasing epidemiological evidence, low vitamin D concentrations have achieved substantial strength as having a key role on MS risk. It is generally thought that the positive effects of raised vitamin D levels may partly be mediated by beneficial increased UVR exposure. In addition, genetic factors were considered as influencing circulating vitamin D levels or modifying the effect of vitamin D, which will be discussed in the section, "The genetics of MS " and Chapters 3 to 5.

1.6.1.1 Sun exposure and the vitamin D synthesis pathway

As described in **Figure 1.2**, the primary form of vitamin D comes from two types of sources: skin exposure to ultraviolet radiation (UVR) in sunlight and diet (e.g., fish and vitamin D supplementation). Pre-vitamin D3 is formed in skin upon photolysis of 7-dehydrocholesterol by UVR, which is then converted to vitamin D3 [67]. In the liver, vitamin D3 is enzymatically converted by 25-hydroxylase to 25-hydroxyvitamin D (25(OH)D), which then undergoes second hydroxylation in the kidney by the enzyme 1α-hydroxylase converting it to the active form of vitamin D, 1,25-dihydroxyvitamin D ($1,25(OH)_2D_3$). It is known that most circulating vitamin D in humans comes from sun exposure [68], and generally 25(OH)D is used for assessment of an individual's vitamin D status, as 25(OH)D is the major circulating form of vitamin D and is not tightly regulated as $1,25(OH)_2D_3$ [69]. 1α-hydroxylase is known to exist in many other tissues,

including immune system and nervous system cells, allowing extra-renal production of the active hormone that is not under the control of the calcium homeostasis system. $1,25(OH)_2D_3$ binds and activates the vitamin D receptor (VDR), a transcription factor that regulates the expression of over 2000 genes [70], which affect calcium and phosphate regulation or influence cells proliferation [71].

Figure 1.2. UV radiation and vitamin D synthesis



1.6.1.2 UVR and MS risk

It has been demonstrated that the incidence, prevalence and mortality of MS typically follows a latitudinal gradient, both in the northern and southern hemispheres [72-76]. In

Europe and North America, MS is more common in the northern regions, while it is more prevalent in the southern part of Australia [77] and New Zealand [78]. These areas, with higher latitude, generally have lower annual sunlight, although genetic effect cannot be ruled out. Research has found individuals with higher residential and/or occupational solar exposure have a lower MS incidence [79,80], and those born in spring had a higher MS risk compared to those born in late autumn and early winter months in both the northern and southern hemisphere [81-85]. The underlying mechanisms may be a protective effect of UVR-induced immunosuppression in the pregnancy possibly mediated directly by maternal UVR or by derived vitamin D [83,86,87]. Experiments conducted using experimental autoimmune encephalomyelitis (EAE), an animal model of CNS inflammation used to study MS, showed UVR can prevent or delay the clinical symptoms of EAE in mice [88]. The association between UVR and MS was initially thought to be solely due to UVR's actions to increase vitamin D levels. However, a potentially independent role of UVR on MS onset has been demonstrated. A review pointed out UVR suppresses the immune system through mechanisms independent of vitamin D [89], and EAE models indicated that UVR suppresses CNS inflammation independent of vitamin D production [90]. In addition, study showed sun exposure and vitamin D are independent risk factors for CNS demyelination, such that higher levels of past, recent, and accumulated leisure-time sun exposure were each associated with a reduced risk of first demyelinating events (FDE) [91].

1.6.1.3 25-hydroxyvitamin D and MS risk

Typically, MS patients have lower serum 25(OH)D levels than healthy controls [88], especially after MS onset [92,93]. Even prior to the onset of MS, 25(OH)D is also considered an important predictor for developing MS in healthy young white adults, independently from their place of birth and latitude of residence during childhood [93].

As described in **Figure 1.2**, 25(OH)D is an integrated measure of vitamin D derived from both UVR exposure and diet. It has been noted in Norway that consumption of fatty seafood and cod liver oil, both rich sources of vitamin D, protected against the risk of MS [94]. In a prospective investigation estimating the association between vitamin D intake and MS risk, researchers found those subjects in the top quintile of vitamin D intake had higher 25(OH)D concentrations (mean 75 nmol/L vs. 55 nmol/L) and a 33% lower incidence of MS, relative to those in the bottom quintile [95]. Interestingly, one study found that the relative risk (RR) of MS was 0.95 (95% CI: 0.88–1.02) for a 10 nmol/l increase in 25(OH)D even after adjustment for UV dose [91], which indicated that 25(OH)D alone may predict subsequent MS risk. In general, 25(OH)D concentrations above 50 nmol/L have been considered adequate, but evidence suggests that a minimum of 75 nmol/L, and perhaps more than 90 nmol/L, is optimum for many health outcomes [96,97]. Increasing outdoor activity and dietary consumption of fish or vitamin D supplements has been recommended to increase the levels of 25(OH)D concentrations; however, genetic factors can also affect 25(OH)D levels.

1.6.1.4 Vitamin D and MS activity and progression

Studies investigating vitamin D for its role in MS clinical course have found levels of 25(OH)D are lower during relapse relative to remission [98-100]. In the prospective Tasmanian MS longitudinal study, an inverse association between seasonal 25(OH)D levels and monthly relapse rates was observed [101], and 25(OH)D inversely associated with relapse risk was demonstrated too [102]. Also, a number of other cohort studies are consistent with their findings of an inverse association between serum 25(OH)D and the risk of relapse, including retrospective [103,104] and prospective cohort designs [105]. These associations with relapse are likely mediated by the immunomodulatory effects of the active metabolite of vitamin D, 1,25-dihydroxyvitamin D ($1,25(OH)_2D_3$), acting to up

regulate regulatory T-cell function and depress inflammatory immune activity [88].

In regards to disease severity, most of the studies demonstrated a significant association with vitamin D levels. A cross-sectional study showed a negative correlation between 25(OH)D levels and disease severity [106]. In the study of 193 MS patients, higher 25(OH)D and 24,25(OH)D were significantly correlated with a decreased MSSS and marginally with decreased EDSS [106]. In addition, an inverse association between 25(OH)D and EDSS were observed in a case-control study [92] and retrospective studies [103,107]. This is probably due to reverse causality where an increased EDSS results in going less outside and result in lower vitamin D levels, and only prospective longitudinal studies can evaluate the causal direction. Interestingly, in the EPIC longitudinal cohort study, vitamin D levels were inversely associated with MS activity on brain MRI, and each 10ng/ml higher 25(OH)D level was associated with lower subsequent disability (β: -0.047; 95% CI: -0.091 to -0.003) [108]. However, research into the association between the active form vitamin D $1,25(OH)_2D_3$ and MS severity as measured by EDSS, showed no association [103].

Overall, the epidemiological evidence demonstrates a reasonably consistent picture that sufficient vitamin D and /or UVR exposure can reduce MS risk, additionally recent literature provided evidence for a causal interpretation. Therefore, intervention with vitamin D supplementation could have a substantial impact on reducing disease burden. Suggesting that over 70% of MS cases in the USA and Europe could be prevented by increasing the serum 25(OH)D concentration of adolescents and young adults to above 100 nmol/L [109,110]. Such high serum 25(OH)D is only common in those with outdoor lifestyles in sunny regions, while alternatively can be reached by taking1000–4000 IU

vitamin D3 daily for most people [111-113]. However, important questions remain about the timing of an intervention, dose-response and the efficacy in slowing progression and effects in different ethnic groups. Currently, not many clinical trials have been established. A small phase I/II study which compared two groups of MS patients on supplementation with either high (up to 40,000 IU/d) or low doses (≤ 4,000 IU/d) of vitamin D3 for 52 weeks showed no significantly increase 25(OH)D levels, though high-dose vitamin D (approximately 10,000 IU/day) in MS is safe [114]. Another small randomised trial of high-dose vitamin D2 (targeting 25OHD 130-175 nM) showed no significant effect in reducing MRI lesions in patients with RRMS compared to low-dose supplementation (1,000 IU/d) [115]. However, a randomised, double blind, placebo controlled trial with 66 MS patients showed vitamin D3 as an add on treatment to interferon β-1b reduces MRI disease activity in MS, though it showed no significant differences in the annual relapse rate compared to placebo add-on group [116]. These varying results indicated that developing well-designed large sample-size clinical trials of vitamin D supplementation in MS are warranted.

1.6.2 Epstein-Barr virus infection

Epstein-Barr virus (EBV) has been demonstrated with consistent evidence as being associated with MS risk. Research into serological exposure to EBV in MS patients found nearly 100% of MS cases with exposure to EBV infection [117-123]. Many case-control studies have found higher frequencies of detection of EBV [122,124-127] and/or higher titres in cases [118-120,125,126,128] relative to controls, although some studies have had inconsistent findings [123,129]. Research found those with an undiagnosed EBV infection in childhood have 10-times greater risk of MS than those not infected with EBV [130]. A systemic review including eight seroepidemiological case-control studies demonstrated an

estimated odds ratio of 13.5 (95% CI, 6.3-31.4) for developing MS for those EBV seropositive compared to those individuals who are EBV seronegative [131]. In addition, consistent findings demonstrated an increase of anti-EBNA antibodies, driven by a marked increase of anti- EB nuclear antigen (EBNA) 1 antibodies and a less prominent increase in anti-EBNA 2 in MS [132-134]. The increase of titers to EBNA complex and EBNA 1 suggests a more severe primary infection or reactivation of infection accompanied by a vigorous cellular immune response [135-137]. A prospective cohort study and a nested case-control study used samples prior to MS onset also support a role of EBV in the aetiology of MS [132,138]. An association between MS and infectious mononucleosis (IM), a manifestation of acute EBV infection commonly seen after puberty has been demonstrated. A meta-analysis including 14 studies indicated that the relative risk of MS after IM was 2.3, such that both in adolescents and young adults, EBV is a risk factor for MS [139]. These findings have also been supported by a more recent systematic review [140].

The pathogenesis of EBV infection on MS is not completely known. Research found T-cells from Cerebrospinal fluid (CSF) and blood of MS patients reacted against EBV-infected B-cells [141], and in a study comparing CD4-positive T-cells from MS patients and controls, EBNA-1-specific memory T-cells had greater frequency, proliferative capacity, and interferon production [142]. These findings suggested that cell-mediated immunity is involved in MS, with a special reactivity against EBV antigens. However, as the direct aetiologic evidence, EBV virus was detected in the brain in some individual studies [143,144], but not all [145,146]. These conflicting studies gave rise to significant discussion, but also indicated that the contribution of EBV to the cause of MS is not yet established, but a relationship is clearly present [147,148].

1.6.3 Smoking

Smoking has been evaluated as a risk factor for MS for decades. Although some case-control or cohort studies showed there were no associations with MS between smokers and nonsmokers [149,150], most studies, especially the larger studies, have demonstrated a consistent association between smoking and MS susceptibility. In the US Nurses' Health Study (121,700 women at baseline in 1976, follow-up to 1994) and Nurses' Health Study II (116,671 women at baseline in 1989, follow-up to 1995), an increased relative incidence rate of MS was observed among current smokers (RR=1.6 (95% CI: 1.2–2.1)) and past smokers (RR= 1.2(95% CI: 0.9–1.6)) compared to those who never smoked, with a clear dose-response effect [151]. Later in 2009, a large, multinational European case–control study including Sweden, Norway and UK showed a consistent result that cigarette smoking was associated with MS risk [152].

Smoking also may exert influence on the course or progression of established MS. It has been demonstrated that smoking increased the risk of conversion from relapsing–remitting to secondary progressive disease, for ever/current smokers compared with never smokers [153,154]. Moreover, ever-smokers were more likely to have progressive disease compared with never-smokers [155,156]; additionally this effect was stronger in those starting smoking early [155] or with heavier cigarette consumption [157]. An increased frequency of relapses [155] and number of lesions visible on MRI in MS patients who smoked compared to nonsmokers was also observed [157].

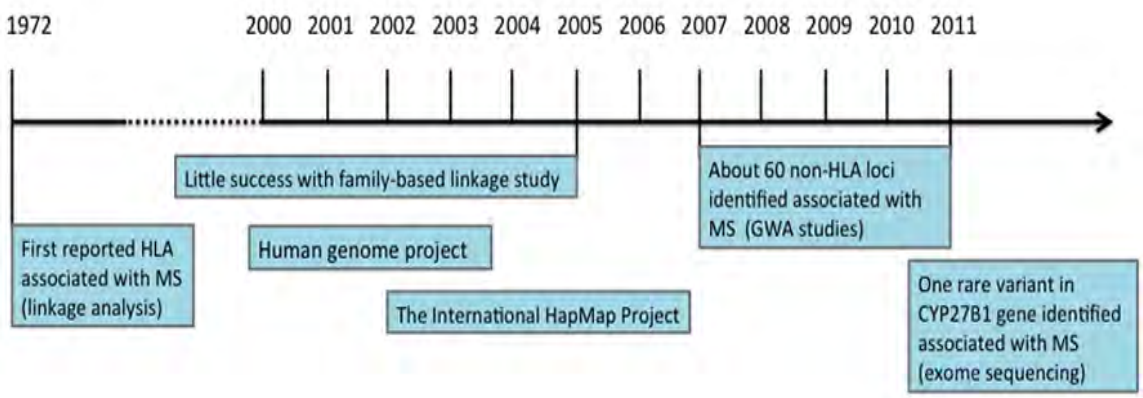
The mechanisms by which smoking might influence the risk of MS and its clinical course are unclear, since over 4,500 compounds and chemicals are contained in tobacco and the resultant smoke from cigarettes, including a number of known carcinogens and other toxic

substances [158]. Among these substances, research has indicated nicotine in the cigarette has immunological effects in T-cell immunosuppression [159]. However, another form of tobacco, snuff use, has not shown an increase in MS risk, which suggested nicotine alone may be unlikely to explain the increased MS risk [152,160]. Despite snuff use, cigarette smoke have been demonstrated to affect numerous immune functions such as T-cell, B-cell, and natural killer cell functions [161-163], which is thought important in MS, whereas it is also associated with other autoimmune diseases, such as systemic lupus erythematosus and rheumatoid arthritis [164,165]. As pointed out by a recent review [166], the overall magnitude of effect of smoking on MS susceptibility or disease course is small, and detecting small effects are more likely to have inconsistent results [167], thus more observational and mechanistic research is needed.

1.7 The genetics of MS

In the 1920s, a worldwide investigation of MS prevalence found that racial and ethnic differences might indicate a genetic influence [168]. Later studies found MS aggregates in families [169], especially the risk to siblings compared to the general population (OR 16.8 (95% CI: 14.0–20.3)) [170]. This was further supported by twin studies showing higher concordance rates in monozygotic twins (24–30%) than dizygotic twins (3–5%) [171,172], and by the observations that there is no increased risk for adoptive relatives [173] and an intermediate risk for half-siblings [174] and offspring of conjugal pairs [175]. Although these epidemiologic data suggest that genetic variation is important for susceptibility to MS, gene discovery has lagged significantly until recently (**Figure 1.3**).

Figure 1.3. The timeline of genetic research in MS.



1.7.1 Limited success for family-based linkage analysis and candidate-gene studies

Family-based linkage analysis has been powerful in identifying the genes responsible for Mendelian diseases whose traits follow monogenic patterns of inheritance, e.g. cystic fibrosis. It has been less helpful in identifying the genes underlying complex traits such as MS. In the 1970s [176,177], linkage analysis showed that MS was associated with the major histocompatibility complex (MHC) region on chromosome 6 encoding the human leukocyte antigen (HLA) genes. Subsequently, several large MS linkage analyses [178-180] have failed to identify any locus outside the MHC region, even though the largest study included 730 multiplex families with 2,692 individuals [181]. The lack of success relates to the limitations of classical genetic techniques: for example, microsatellite-based screens and low density marker sets, along with disease heterogeneity and the absence of large extended families with clear and homogeneous mode of transmission. Candidate-gene studies are the practical alternative to linkage analysis, and these have successfully identified some genes contributing to susceptibility of common diseases, but not MS. Generally, candidate-gene studies are based upon biological hypotheses or knowledge of the candidate-gene's location through linkage analysis. When the fundamental

physiological basis of a disease is unknown, as in MS, the candidate-gene approach is usually unsuccessful.

1.7.2 Success with Genome Wide Association Studies (GWASs)

The HapMap project [182] defined linkage disequilibrium patterns across the human genome, allowing development of GWAS tools that use single nucleotide polymorphisms (SNPs) as markers of genetic diversity directly on the basis of linkage disequilibrium. The HapMap project defined roughly 11 million common genetic variations (SNPs) in the human genome and found that many groups of neighbouring SNPs correlate nearly perfectly with each other. This allowed the selection of variants that represent short regions of the genome [183]. Many SNPs target a genetic locus and may indicate associations with many genes in an area of linkage disequilibrium within the genome. Therefore, it is important to take care when assigning a SNP to a specific gene. The HapMap data allowed the development of high-density SNP mapping of the genome with technologies that simultaneously assess over a million SNPs on a single chip in a highly cost-effective manner.

The major problem with GWASs is the statistical burden of multiple testing, such that larger and larger studies are needed to detect variants with smaller and smaller odds ratios. Despite this, GWASs have been highly successful in finding common variants in many complex disorders. Also, newly developed statistical methods have improved the analysis, such as the identifying and correcting for population stratification and relatedness [184,185], and imputating ungenotyped variants [186]; these approaches increase the accuracy and reliability of GWASs outputs.

To date, GWASs have detected over 100 loci associated with MS outside the HLA region [187], and have confirmed the major role of the *HLA-DRB1*15:01* (*HLA-DR15*) gene (OR 3.08, $p<10^{-312}$) [188]. All the non-HLA associations are common variants with modest risk (OR in the region of 1.1–1.3), and many are near genes with key roles in the immune system (**Table 1.1**). A large MS GWAS conducted by the International Multiple Sclerosis Genetics Consortium (IMSGC) and Wellcome Trust Case Control Consortium 2 (WTCCC2) studied 9,772 cases and 17,376 controls and identified a further 29 novel suscepbility loci, while replicating almost all of the previously associated loci (**Table 1.1**). It also pointed out that the identified genetic regions were in, or near to, genes with a primary role in T-cell-mediated immune mechanisms [188].

Table 1.1. Non-HLA loci identified from GWAS with strong evidence conferring susceptibility for MS

SNP (first reported)	Chr	Position	MAF	OR	P-value		Nearest gene	Gene function*	Pathway#
					First reported	Latest reported			
rs3748816	1	2,516,606	0.70	1.16	3.54E-06[187]	1.10E-13[186]	*MMEL1*	Neuropeptide degradation, degradation of beta-amyloid.	Neurological relevance
rs6680578	1	92,949,466	0.38	1.11	5.0E-04[188]	6.50E-12[186]	*EVI5*	Cell cycle; cell proliferation; cell division; retroviral integration in T-cell lymphomas	Immune system
rs6604026	1	93,076,191	0.29	1.14	7.94E-06[188]	2.5E-06 [189]	*RPL5*	For rRNA maturation and formation of the 60S ribosomal subunits	N/A
rs11581062	1	101,180,107	0.71	1.13	2.50E-10[186]	N/A	*VCAM1*	Mediates leukocyte-endothelial cell adhesion and signal transduction	Immune relevance
rs12044852	1	116,889,302	0.08	1.24	1.9E-05[188]	2.00E-09[186]	*CD58*	Cell adhesion, T-cell receptor signaling; T-lymphocyte activation	Immune relevance (co-stimulatory)
rs2760524	1	190,797,171	0.16	0.87	9.77 E-06[190]	8.80E-07[186]	*RGS1*	Activates G-proteins	Immune relevance (signal transduction)
rs12122721	1	199,251,103	0.28	1.22	6.56E-10[191]	4.10E-07[186]	*KIF21B*	Neuronal (axonal) transport	Neurological relevance
rs6718520	2	43,179,074	0.48	1.17	3.42E-08[192]	N/A	*THADA*	Thyroid adenoma-associated	N/A
rs7595037	2	68,500,599	0.55	1.1	5.10E-11[186]	N/A	*PLEK*	PLEK protein variant	N/A
rs17174870	2	112,381,672	0.25	1.1	1.30E-08[186]	N/A	*MERTK*	Cell entry factor	N/A
rs882300	2	136,692,725	0.46	0.84	1.37E-07[190]	N/A	*CXCR4*	Acts with the CD4 protein	Immune system
rs10201872	2	230,814,968	0.185	1.13	1.80E-10[186]	N/A	*SP140*	Involved in the pathogenesis of acute promyelocytic leukaemia and viral infection.	N/A
rs11129295	3	27,763,784	0.36	1.11	1.20E-09[186]	N/A	*EOMES*	Trophoblastic development and gastrulation	N/A
rs170934	3	28,054,089	0.48	1.17	1.57E-08[192]	N/A	*EOMES*	Trophoblastic development and gastrulation	N/A
rs9657904	3	107,069,404	0.24	1.40	1.6E-10[193]	2.10E-04[186]	*CBLB*	Encodes a negative regulator of adaptive immune responses	I Immune relevance (signal transduction)
rs1132200	3	120,633,526	0.16	1.24	3.09E-9[191]	1.10E-07[186]	*TMEM39A*	N/A	N/A
rs9282641	3	123,279,458	0.09	1.21	1.00E-11[186]	N/A	*CD86*	Activation of the T-cell; diminishes the immune response	Immune relevance (co-stimulatory)
rs4680534	3	161,181,639	0.37	1.12	5.58 E-06[190]	3.70E-06[186]	*IL12A*	Acts on T and natural killer cells	Immune relevance (cytokine pathway)
rs228614	4	103,797,685	0.45	1.09	1.40E-07[186]	N/A	*NFKB1*	A pleiotropic transcription factor, involved in many biological processed such as inflammation, immunity, differentiation, cell growth, tumourigenesis and apoptosis.	Immune relevance
rs6897932	5	35,910,332	0.25	1.18	2.94E-07[188]	1.30E-06[186]	*IL7R*	Lymphocyte development; T-cell receptor gamma accessibility by STAT5; apoptosis; activation of T lymphocytes	Immune relevance (cytokine pathway)
rs6896969	5	40,460,183	0.38	0.91	2.40 E-07[190]	6.90E-14[186]	*PTGER4*	Activates T-cell factor signaling	Immune relevance
rs2546890	5	158,692,478	0.52	1.1	1.20E-11[186]	N/A	*IL12B*	As a growth factor for activated T and NK cells,	Immune relevance

								enhance the lytic activity of NK/lymphokine-activated killer cells, and stimulate the production of IFN-gamma	(cytokine pathway)
rs12212193	6	91,053,490	0.53	1.09	3.80E-08[186]	N/A	*BACH2*	Coordinate transcription activation and repression	N/A
rs802734	6	128,320,491	0.68	1.1	5.50E-09[186]	N/A	*THEMIS*	Involves T-cell selection during late thymocyte development	Immune relevance
rs11154801	6	135,781,048	0.36	1.15	1.00E-13[186]	N/A	*MYB*	Transcriptional activator; control proliferation and differentiation of hematopoietic progenitor cells	N/A
rs17066096	6	137,494,601	0.76	1.14	6.00E-13[186]	N/A	*IL22RA2*	An IL22 antagonist in the regulation of inflammatory responses	Immune relevance (cytokine pathway)
rs9321619	6	137,916,101	0.47	0.89	1.71 E-05[190]	2.30E-06[186]	*OLIG3*	Determine the distinct specification program of class A neurons in the dorsal part of the spinal cord and suppress specification of class B neurons.	N/A
rs1738074	6	159,385,965	0.43	1.13	6.80E-15[186]	N/A	*TAGAP*	A GTPase-activating protein; may play important roles during T-cell activation	Immune relevance (signal transduction)
rs354033	7	148,920,397	0.26	1.1	4.70E-09[186]	N/A	*ZNF767*	Zinc finger family member 767	N/A
rs4410871	8	128,884,211	0.28	1.11	7.70E-09[186]	N/A	*MYC*	Participates in the regulation of gene transcription	N/A
rs2019960	8	129,261,453	0.77	1.1	5.20E-09[186]	N/A	*PVT1*	N/A	N/A
rs2150702	9	5,883,861	0.49	1.16	3.28E-08[192]	N/A	*MLANA*	Involved in melanosome biogenesis	N/A
rs12722489	10	6,142,018	0.15	1.25	2.96E-08[188]	2.00E-09[186]	*IL2RA*	Elimination of self-reactive T-cells	Immune relevance (cytokine pathway)
rs1250540	10	80,706,013	0.35	1.12	1.59 E-06[190]	1.40E-06[186]	*ZMIZ1*	Increases ligand-dependent transcriptional activity of AR and promotes AR sumoylation.	N/A
rs7923837	10	94,471,897	0.38	1.1	4.90E-09[186]	N/A	*HHEX*	Transcriptional repressor; may play a role in hematopoietic differentiation	N/A
rs17824933	11	60,517,188	0.25	1.18	3.79E-09[190]	1.70E-09[186]	*CD6*	Cell adhesion; T-cell activation	Immune relevance
rs630923	11	118,259,563	0.15	1.11	2.80E-07[186]	N/A	*CXCR5*	Involves B-cell migration	Immune relevance (cytokine pathway)
rs1800693	12	6,310,270	0.45	1.20	1.59E-11[190]	9.20E-11[186]	*TNFRSF1A*	Apoptosis; cytokine signaling	Immune relevance (cytokine pathway)
rs10466829	12	9,767,358	0.50	1.09	1.40E-08[186]	N/A	*CLECL1*	May play a role in the interaction of dendritic cells with T-cells and the cells of the adaptive immune response.	Immune relevance (co-stimulatory)
rs703842	12	56,449,006	0.29	0.81	5.4E-11[189]	2.40E-06[186]	*CYP27B1*	Catalyzes the conversion of 25-hydroxyvitamin D3 (25(OH)D) to 1-alpha,25-dihydroxyvitamin D3 (1,25(OH)2D) plays an important role in normal bone growth, calcium metabolism, and tissue differentiation	Vitamin D
rs1790100	12	122,222,678	0.24	1.11	7.21 E-07[190]	1.50E-04[186]	*MPHOSPH9*	M-phase phosphoprotein 9	N/A
rs9523762	13	92,129,887	0.35	1.36	7.0E-06[194]	N/A	*GPC5*	Cell surface proteoglycan	N/A
rs4902647	14	68,323,944	0.47	1.11	9.30E-12[186]	N/A	*ZFP36L1*	Regulates the response to growth factors	N/A
rs2300603	14	75,075,310	0.74	1.11	2.00E-08[186]	N/A	*BATF*	A negative regulator of AP-1 mediated transcription	N/A
rs2119704	14	87,557,442	0.08	1.26	2.20E-10[186]	N/A	*GALC*	Hydrolyzes the galactose ester bonds	Neurological relevance
rs2744148	16	1,013,553	0.83	1.12	8.40E-08[186]	N/A	*SOX8*	May play a role in CNS, limb and facial development.	Neurological relevant
rs6498169	16	11,156,830	0.37	1.14	3.83E-05[188]	6.30E-14[186]	*CLEC16A*	Glycoprotein uptake for dendritic cell presentation; dendritic cell distinguishing between self and non-self antigens	Immune relevance
rs17445836	16	84,575,164	0.19	1.25	3.73E-09[190]	7.00E-08[186]	*IRF8*	Regulation of immune cells	Immune relevance (cytokine pathway)
rs744166	17	37,767,727	0.56	0.87	2.75E-10[195]	3.50E-06[186]	*STAT3*	Transcription factor that binds to the interleukin-6 (IL-	Immune relevance

								6)-responsive elements	(signal transduction)
rs180515	17	55,379,057	0.65	1.11	8.80E-08[186]	N/A	*RPS6KB1*	Phosphorylates specifically ribosomal protein S6 in response to insulin or several classes of mitogens.	N/A
rs7238078	18	54,535,172	0.77	1.11	2.50E-09[186]	N/A	*MALT1*	Enhances BCL10-induced activation of NF-kappa-B	Immune relevance (signal transduction)
rs763361	18	150,069,017	0.48	1.13	4.54E-04[196]	N/A	*CD226*	Tumour suppression; T-cell differentiation and proliferation; monocyte migration; antigen presentation; cell adhesion	Immune relevance
rs1077667	19	6,619,972	0.21	1.16	9.40E-14[186]	N/A	*TNFSF14*	Activates NFKB, stimulates the proliferation of T-cells.	Immune relevance (cytokine pathway)
rs34536443	19	10,324,118	0.047	1.32	2.7E-06[197]	1.50E-06[186]	*TYK2*	Involved in the initiation of type I IFN signaling	Immune relevance (signal transduction)
rs874628	19	18,165,700	0.72	1.12	1.30E-08[186]	N/A	*MPV17L2*	N/A	N/A
rs2303759	19	54,560,863	0.75	1.11	5.20E-09[186]	N/A	*DKKL1*	Interacts with the Wnt signaling pathway	N/A
rs6074022	20	44,173,603	0.28	1.20	1.3E-07[189]	1.70E-06[186]	*CD40*	Stimulatory molecule on B-cells, dendritic cells, macrophages and microglia.	Immune relevance (co-stimulatory)
rs2248359	20	52,224,925	0.39	1.12	2.50E-11[186]	N/A	*CYP24A1*	Regulates the level of vitamin D3	Vitamin D
rs6062314	20	61,880,157	0.92	1.17	1.30E-07[186]	N/A	*TNFRSF6B*	Decoy receptor for the cytotoxic ligands TNFS14/LIGHT and TNFSF6/FASL; protects against apoptosis.	N/A
rs2283792	22	20,461,125	0.48	1.09	4.70E-09[186]	N/A	*MAPK1*	Involved in both the initiation and regulation of meiosis, mitosis, and post-mitotic functions in differentiated cells	N/A
rs140522	22	49,318,132	0.33	1.09	1.70E-08[186]	N/A	*SCO2*	Acts as a copper chaperone, transporting copper to the Cu(A) site on the cytochrome c oxidase subunit II (COX2).	N/A

Note 1.1 Association between HLA and MS identified from GWASs.

The information about the HLA region that was identified as associated with MS in GWASs was not provided in our original published paper. It is necessary to provide the relevant data, since HLA region by far is the strongest region identified to be associated with MS. **Note 1.1 Table A** presents the available data reported by Sawcer and colleagues [188]. Both class I (*HLA-A*) and class II (*HLA-DRB1*) in the HLA region have been confirmed as protective against or increasing risk of disease onset. Of note, although high linkage disequilibrium (LD) across the HLA region, the predominant association with MS susceptibility is with the class II *DRB1* haplotype (*DRB1*15:01*, *DRB1*13:03*, *DRB1*3:01*, *DRB1*08:01*), and the major role is *DRB1*15:01* with an OR of 3.08 ($p<10^{-312}$) [188].

Note 1.1 Table A. HLA loci identified as associated with MS by GWASs

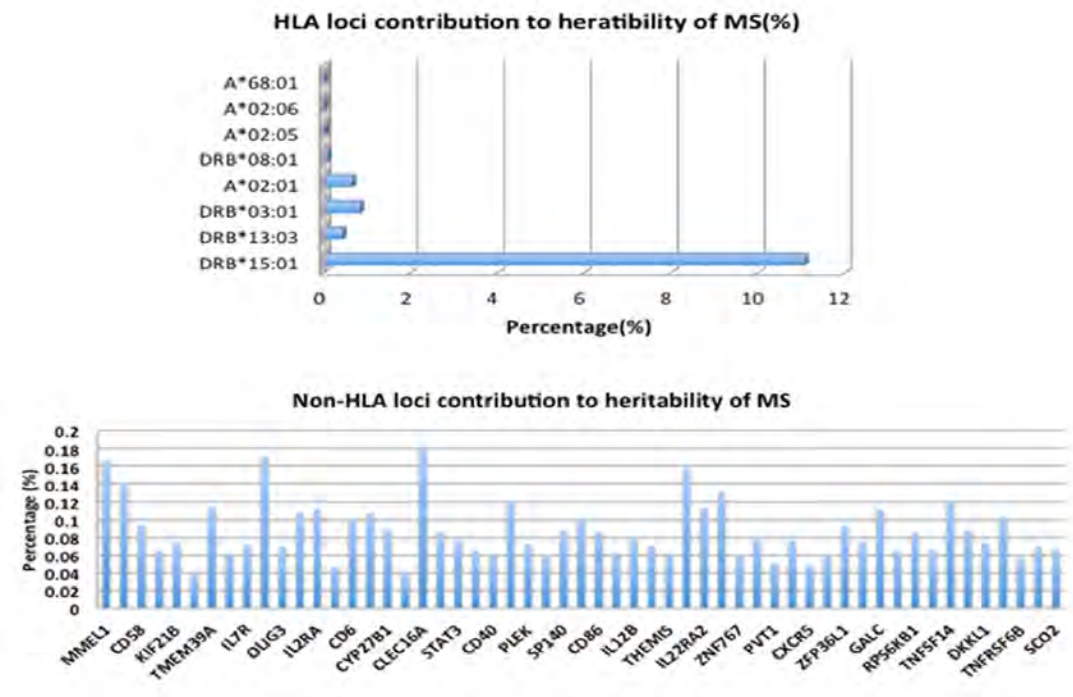
Locus/allele	OR	p-value
DRB*15:01	3.08	<1E-312
DRB*13:03	2.43	1.30E-11
DRB*08:01	1.18	1.60E-07
DRB*03:01	1.26	3.60E-10
A*02:01	0.73	9.10E-23
A*02:05	0.76	1.50E-01
A*02:06	1.23	5.10E-01
A*68:01	0.97	6.10E-01

1.7.3 GWASs explain a small part of the heritability of MS

Although GWASs have identified over 60 loci associated with MS by 2012, these explain only a small fraction of MS heritability. In our own studies, taking account of the

identified loci, allele frequencies and sibling relative risk using Risch's gene effect measuring method [200], we found all the identified loci explained only 18–24% of the heritability of MS [170]; *HLA-DRB1*15:01* explained 11% (**Figure 1.4**). Hence the obvious question: "Where is the missing heritability?"

Figure 1.4. Genetic contribution to MS risk. In total, 18% of heritability of MS was explained by the loci identified in the biggest MS GWAS[188]. 8 HLA-loci explain 13.2% of heritability of which *HLA-DR15* explains 11.1%; 57 non-HLA loci explain 4.8% of MS heritability with *CLEC16A* the largest contributor at 0.18% of heritability.



1.7.4 Where is the missing heritability in MS?

Several as yet unproven theories may explain the missing heritability:

1) **Rare variants**. GWAS are constrained by their design to the study of common variants (minor allele frequencies (MAF) ≥5%); it is hard to detect rare variants (MAF <5%) by standard genotyping arrays.

2) **Common variants**. There are probably many more undiscovered common variants with even smaller effects (1<OR<1.1) waiting to be found by larger and larger GWAS. However, one recent modeling exercise in Crohn's disease suggested that even for an infinitely-sized GWAS, common variants explained less than 50% of heritability [201].

3) **Epigenetics**. Epigenetic modifications may regulate the expression or activity of genes; the effect of a gene may increase by either increasing or decreasing its expression.

4) **Gene–gene interactions**. If two or more genes govern a single phenotype, the genes will affect each other's expression and therefore change their effect.

5) **Pathway involvement**, where strongly associated variants cluster with weak ones within a known signaling pathway; this may increase the overall effect of the pathway and enhance the genetic susceptibility to disease.

6) **Gene–environment interaction**, where the environment increases or decreases the effect of a gene in determining disease onset. For example, the genetic defect underlying phenylketonuria, although completely penetrant when there are two copies of the abnormal gene, is phenotypically silent unless the environmental factor—dietary phenylalanine—is introduced.

7) **Structural variants** refer to regions of DNA from around 1Kb to several Mb, and include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as 'copy number variants'. Such variation has been largely unexplored in relation to complex traits [202].

1.7.4.1 Rare Variants

Recently, rare variants have been receiving increased attention in an attempt to explain the 'missing heritability' fraction for many complex disorders including MS. Rare variants with lower MAF (less than 5%) have not been represented on many of the microarrays used for major genetic studies up until very recently and they are often difficult to detect in case-control study designs unless the samples are from an isolated population or have some other enrichment factor allowing detectable numbers of copies for the minor alleles. It is generally thought that rare variants are likely to have larger effect sizes (within the pedigrees or populations in which they segregate) and could contribute significantly to missing heritability. It is postulated that these variants are also likely to have more obvious functional consequences [203]. It is becoming evident that rare variants have a large cumulative effect on normal phenotypic variation and are extremely important in disease [204]. Recently, studies using exome sequencing in a set of multiplex multiple sclerosis families found several rare functional variants in *CYP27B1* underlying the known MS locus on chromosome 12 [205]; this was the first description of a rare variant in MS. Interestingly, *CYP27B1* encodes the protein that converts 25-hydroxyvitamin D3 to the active hormone 1,25-dihydroxyvitamin D3. The rare functional variants in *CYP27B1* are associated with the development of vitamin D-dependent rickets. The authors noted that in all proven genetic cases of vitamin D-dependent rickets in Norway, there was a 100% concordance with MS. We await further evaluation of this finding in other cohorts. This finding also supports the increasing evidence for a role for vitamin D deficiency in the onset of MS.

Note 1.2 Evidence in disease aetiology and detecting strategy for rare variants

In Chapter 2 and Chapter 3, we describe how to detect rare variants and/or disease

causing variants that confer susceptibility to MS. Thus, we here added here some more background of rare variants to better understand the two chapters.

It is known that most SNPs in the human genome are rare SNPs (**Figure note 1.1**) [206] according to the Encyclopedia of DNA Elements (ENCODE) and the International HapMap Database (http://hapmap.ncbi.nlm.nih.gov/). Rare variants are likely to have larger effect sizes (within the pedigrees or populations in which they segregate) and could contribute significantly to missing heritability. Based on the published results [207], most identified rare variants have larger effect size (OR>2) than those of common variants (OR: 1.1-1.3). More recently, four rare variants within the *IFIH1* gene were identified to have a larger individual effect on the risk of type I diabetes than those of common variants [208]. For the *BRCA1* and *BRCA2* genes, many rare variants in these genes are associated with an increased risk of breast cancer by an average odds ratio of ~10 in randomly-ascertained cases aged 60-70 [209]. Rare variants are also likely to have more obvious functional consequences [203]. There is rapidly accumulating data showing that rare variants have a large cumulative effect on normal phenotypic variation and are extremely important in disease [204].

Figure note 1.1. Distribution of single-nucleotide polymorphisms (SNPs) from the Encyclopedia of DNA Elements (ENCODE) (orange bars) and of all SNPs reported in the International HapMap Database (blue bars) by minor allele frequency (MAF).

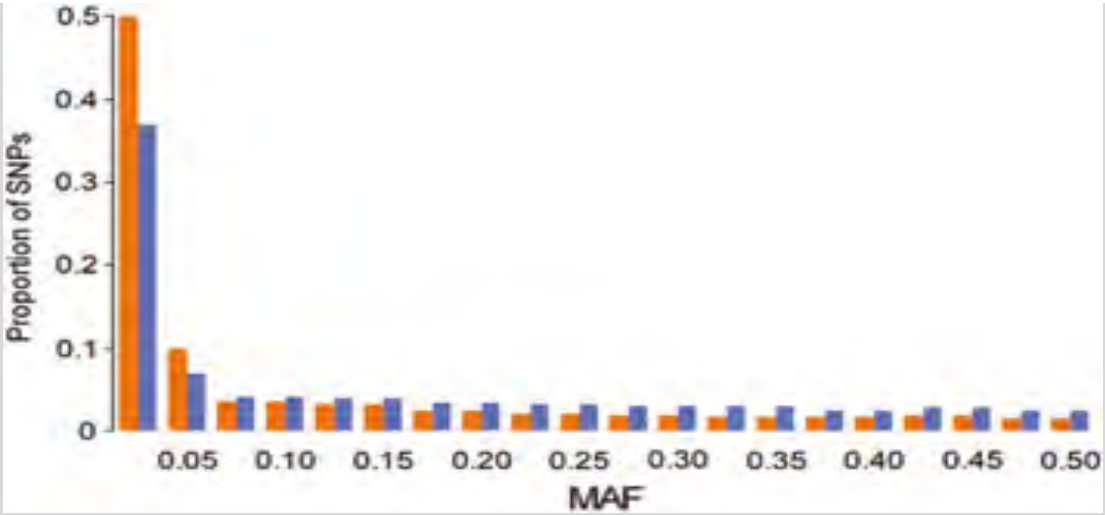


Figure reproduced from Gorlov and colleagues [206]

Note 1.2.1 Evidence of rare variants in disease aetiology

Empirical evidence has already indicated that rare SNPs may be crucial in the genetic architecture of common diseases. Although scientists are convinced that mutation-selection balance, genetic drift, and effective population size play an important role in shaping the frequency and distribution of genetic polymorphisms in the human genome [210-213], it is not feasible to estimate directly the effects of millions of SNPs. That the functional SNPs should somehow disturb a normal function is to be expected to be under the pressure of selection (mostly negative selection). A stronger effect of a SNP on gene function may be associated with stronger negative selection against it. Thus it is possible to estimate risk on the basis of quantitation of selection pressure. PolyPhen, a common server, can be used to study the selective pressure on the structural and functional properties of proteins [214], which uses empirically derived rules to predict whether a nonsynonymous SNPs (nsSNP) is damaging or benign [215]. Therefore, PolyPhen can also be used to evaluate whether the reported/identified association indeed has a functional meaning, as well as to distinguish causal from non- causal relationships

between a nonsynonymous SNPs and the phenotype of interest [214]. By applying PolyPhen, Gorlov and colleagues analysed the relationship between the MAF and the proportion of nsSNPs predicted to be protein changing. They found a statistically significant negative association between the MAF and the proportion of SNPs predicted to be functional ($p<1^{E-06}$) (**Figure note 1.2**) [216]. A higher proportion of rare nsSNPs are predicted to be functional, and hence rare SNPs are more likely to change disease risk and confer a larger odds ratio. Gorlov and colleagues also evaluated the relatioship between MAF and the change in accessible surface propensity(dprop), one of the most important predictors of SNP functionality in PolyPhen [214], and found there was a negative association between the MAF and the degree of change to the protein structure: rarer SNPs more strongly impair protein structure than do the common variants, which indicates rare SNPs are more likely have a larger effect size than common SNPs [206].

Figure note 1.2. Proportion of nonsynonymous single-nucleotide polymorphisms (nsSNPs) predicted to be protein damaging by minor allele frequency (MAF). Each point represents the proportion of functional nsSNPs in a given MAF category. Error bars indicate standard error. Proportion predicted using the PolyPhen method. The black line is the unadjusted logarithmic regression curve, and the orange line is the curve adjusted for PolyPhen's sensitivity and specificity.

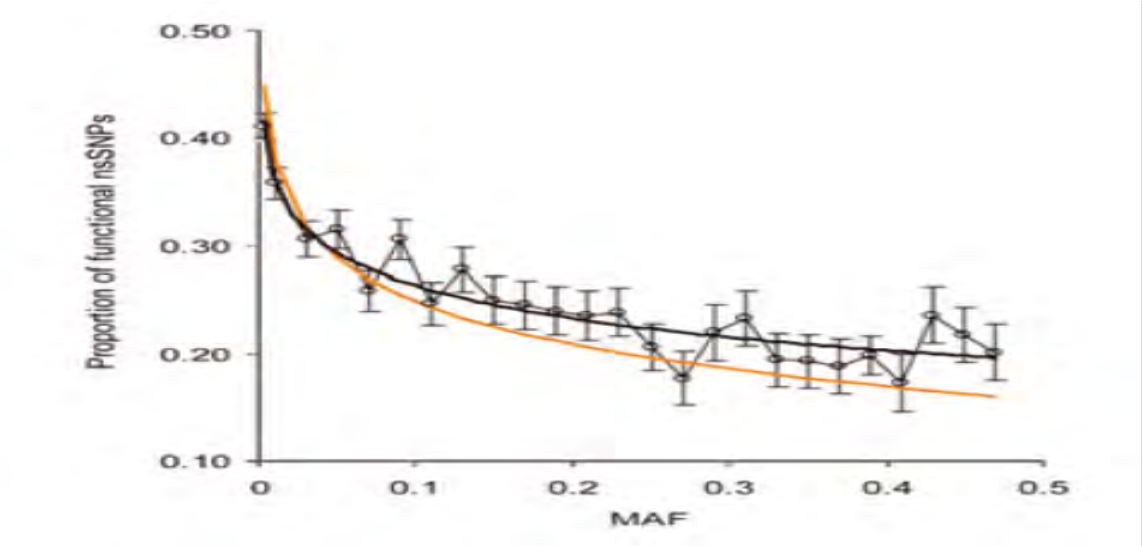


Figure reproduced from Gorlov and colleagues [216].

Note 1.2.2 Strategy for detecting rare variants

Since rare variants may play an important role in disease aetiology, this raises the question: how to detect rare variants? Three strategies have been described and implemented in recent research. One is to re-sequence loci containing common susceptibility variants identified from GWASs. The target genes, by using this strategy, usually contain common disease-associated polymorphisms or mutations with biological or aetiological evidence in prior studies or animal models. For example, four rare variants within the *IFIH1* gene were identified to protect against type 1 diabetes (T1D) disease by re-sequenced exons and splice sites [208] of ten candidate genes which were detected as associated with T1D in a GWAS [217] or in functional [218] and rat model studies [219]. However, this strategy precludes the identification of new loci; and in many instances, the SNPs showing the most significant disease association map to genomic regions with no obvious function, thus re-sequencing would provide few clues as to how causal variants affect the disease gene. The second strategy is exome sequencing or whole-genome

sequencing. Exome sequencing is a robust approach to sequencing the complete coding region (exome) for identification of functional variants that are responsible for diseases onset without the high costs of whole-genome sequencing while maintaining high coverage in sequence depth [220]. Exome sequencing is a cheaper and effective alternative to whole genome sequencing but without the capability to identify the structural and non-coding variants associated with the disease. Eventually it will be possible to overcome this limitation by whole-genome sequencing, but it remains prohibitively expensive to perform adequately powered studies. The third way is to re-analyse GWAS data using 'population-based linkage analysis' (PBLA). PBLA is a linkage analysis applied at the population level to detect megabase-scale regions where many pairs of cases have inherited long haplotypes from distant ancestors, perhaps ten to one hundred generations ago [184,221]. It tests for correlation between identical by descent (IBD) sharing and phenotypic sharing. If a linkage region with IBD was identified, the follow-up targeted re-sequencing allows recovery of a more complete inventory of sequence variation, and enables systematic fine-mapping efforts to identify those putatively causal variants with the strongest effects on disease susceptibility cost effectively. Even when whole-genome sequencing becomes cheap enough to pursue with substantial sample sizes, PBLA may still help reduce the massive multiple-testing problem by prioritising regions.

More recently, a study using exome sequencing in a MS family identified one rare variant in *TYK2* that had a modest effect on MS risk [222]. In a functional study, rs34536443 in *TYK2* has been demonstrated to cause an amino acid substitution in the protein that enhanced activation of the kinase function upon stimulation with interferon-β, IL-6 and IL-10 [223]. This effect on MS susceptibility, as the author pointed out, might be

mediated by deviating T lymphocyte differentiation toward a Th2 phenotype, which is also important to other autoimmune diseases [222].

1.7.4.2 Epigenetics and MS

'Epigenetic' refers to the modification of gene expression or biological phenotypes resulting from changes to the genomic structure that do not alter the underlying genetic code [224]. The major epigenetic processes in mammalian cells are methylation and modification of histones. DNA methylation may be responsible for the stable maintenance of a particular gene expression pattern through mitotic cell division [225]. DNA methylation is generally associated with reduced or suppressed gene expression. Cells acquire epigenetic modifications over time and the methylation signatures persist through cell division. Histone modifications are the post-translational covalent addition of molecules to the histone subunits of the nucleosome [226]. Histone modification acts in diverse biological processes, such as gene regulation, DNA repair, chromosome condensation (mitosis) and spermatogenesis (meiosis). Histone modification may also interact with DNA methylation to regulate gene expression, since specific patterns of histone modifications associate with the pattern of DNA methylation [227].

The role of epigenetics in the aetiology of MS is still unclear. However, several studies have suggested that epigenetic modifications may be important in determining MS susceptibility. Moreover, older MZ twins with MS show significantly different patterns of genomic distribution of 5-methylcytosine DNA and histone acetylation [228], indicating the effect of the external environment on the epigenome [229]. Also, MS shows maternal parent-of-origin effects associated with DNA methylation [230] and these influence the age of MS onset [231]. *HLA-DR15* status also influences the maternal parent-of-origin

effect: thus, MS patients who are *HLA-DR15* positive have an increased female to male ratio and mothers who are *HLA-DR15* positive are more likely to have affected female offspring than those who are *HLA-DR15* negative [232]. However, this effect has not been seen in all studies [233]. Recently, a whole-genome sequencing study, assessing DNA methylation and gene expression in three MS-discordant MZ twin pairs, found no consistent differences in DNA sequence, DNA methylation or gene expression in CD4+ T-cells [234]. Researchers have subsequently argued that these data cannot conclusively rule out possible involvement of somatic changes in twins discordant for MS because of design limitations with the study, particularly the small sample size, the low average depth coverage of the genome sequencing and that the analysis limited to just $CD4_+$ T-cells [235].

1.7.4.3 Gene–gene interactions and MS

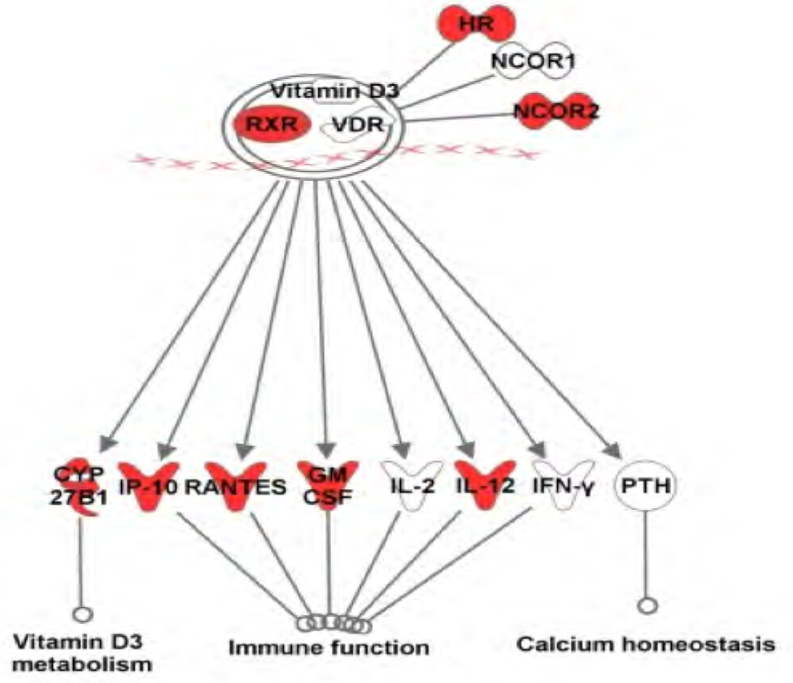
For common diseases driven by complex genetic susceptibility, gene–gene interactions appear to be a potential explanation for the missing heritability [203]. Gene–gene interaction refers to the situation where a single characteristic (phenotype) is governed by two or more genes and each gene affects the expression of the other genes involved. Several studies have examined the role of gene–gene interaction in MS, particularly HLA gene–gene interactions. For example, in a Swedish population, the combination of *HLA-DR15* carriage and absence of *HLA-A*02* alleles increased the risk of MS by 23-fold [236]. However, the largest MS GWAS conducted to date found no definitive evidence for gene–gene interaction in MS susceptibility [188].

1.7.4.4 Pathway analysis and MS

In biological systems, genes and their products (usually proteins) operate within pathways that exert various functional outcomes within the body, such as cell signaling cascades or the metabolism of compounds such as vitamins or amino acids. Pathway analysis involves investigating genes only within a particular pathway and thereby provides specific information about the genetic pathway being affected in disease conditions—yielding biological insights into the disease process. However, most common diseases are influenced by the combined effects of many loci and/or a number of rare variants that interact with environmental factors in multiple complicated ways [207].

Pathway-based analytical approaches allow the genetic variants associated with disease to be considered within the context of a known biological pathway. Evidence of enrichment of susceptibility variants within a pathway may offer more insight into the disease process. If disease-associated variants cluster within a known signaling pathway, then disruption or alteration of that pathway may be involved in the disease process. For example, we considered SNPs shown by GWASs to be associated with MS [191] and found evidence of enrichment of disease susceptibility SNPs within a module of the vitamin D metabolism and signaling pathway (**Figure 1.5)**. This pathway includes known MS-associated genes *CYP27B1* and *IL12* [188]. While some of the remaining SNPs are only weakly associated with MS (we used the top 25% of the GWASs sample) the additional evidence from the pathway analysis, and the postulated involvement of vitamin D in the disease process, strengthens their appeal for further consideration.

Figure 1.5. One module of vitamin D pathway analysis including 13 genes, of which 8 genes have been associated with MS (in red).



To date, there have been few published pathway analyses in MS. One pathway-orientated MS study that took into account all SNPs with nominal evidence of association ($p<0.05$), rather than just those SNPs that exceed the genome-wide significance threshold [237], found three immunological pathways that were over-represented in MS. These included cell adhesion pathways, communication and signaling pathways, and one neural pathway, namely axon-guidance and synaptic potentiation. Recently, by analysing the T-helper cell differentiation pathway, the IMSGC/WTCCC2 GWAS found over-representation of genes involved in T-cell maturation, including 12 genes coding for the cytokine pathway, 6 genes coding for co-stimulatory molecules and 7 genes coding for signal transduction molecules, all of which are immune relevant. These findings indicate that T-cell signaling pathways may be important in the pathogenesis of MS [188].

1.7.4.5 Gene-environment interaction and MS

The known environmental risk factors for MS include high Epstein–Barr virus IgG antibody titres, low vitamin D level and/or UVR exposure and cigarette smoking. Interactions between genes and environmental factors may explain at least part of the missing heritability: a particular genetic variable in the right environmental situation may exert a significantly greater effect on MS causation than if the environmental exposure were not present. So far, attention has largely focused on the major genetic risk factor *HLA-DR15*. Thus, the effect of an environmental exposure on MS might depend on whether a person is *HLA-DR15* positive or *HLA-DR15* negative; conversely, the effect of *HLA-DR15* might depend on the level of the environmental factor.

Since *HLA-DR15* may be important in determining the CD4 Th1-mediated immune response to EBV infections [238], it is possible that EBV and *HLA-DR15* interact with each other. Indeed, several HLA alleles recognize EBNA-1 epitopes [239]. Several epidemiological MS studies have looked at the interaction between *HLA-DR15* and EBV (reviewed in [240]), but so far without any clear association [150].

Several genetic vitamin D pathway variants (e.g. vitamin D receptor (*VDR*) [241], *CYP24A1* [188] and *CYP27B1* [191]) appear to increase MS risk, but few studies have examined interactions with 25(OH)D levels, or reported sun exposure or vitamin D intake. Regarding the vitamin D receptor (*VDR*) gene, the Tasmanian MS case-control study showed that only people with a copy of G at the *Cdx-2* polymorphism (rs11574010) in the VDR gene showed increased MS risk from low winter childhood sun exposure [242], while in the US Nurses Cohorts study, only those with a ‘ff’ genotype of the *FokI*

polymorphism (rs10735810) within the VDR gene showed an effect on MS risk from low vitamin D intake. There was no significant interaction between latitude of residence and *FokI* genotype [243]. The SNPs for *FokI* and *Cdx-2* are not in linkage disequilibrium with one another. There were also no interactions between low vitamin D intake and *CYP27B1, CYP24A1, CYP2R1, GC or HLA-DR15* [243].

There is some interesting functional evidence of an interaction between HLA and vitamin D. Recently, a vitamin D response element (VDRE) was found in the *HLA-DRB1* promoter region[244]. The VDRE was highly conserved in the major MS-associated haplotype *HLA-DR15*, but not conserved among non-MS associated haplotypes [244]. This study also showed that the VDRE responded to $1,25(OH)_2D_3$ and influenced gene expression in B-cells that were transiently transfected with the *HLA-DR15* promoter [244]. These functional studies support the case that higher vitamin D levels at critical times—specifically in *HLA-DR15* positive individuals—might reduce the risk of MS onset. This concept may be in line with the finding that people with MS born in April were more often *HLA-DR15* positive whereas those born in November were less often positive for this gene [245].

Note 1.3 Additional studies on gene-vitamin D interaction influencing MS clinical course

More recently, our prospective cohort study revealed that gene-vitamin D interactions influencing MS clinical course [246,247], including those genes involved in the vitamin D metabolism and VDR/RXR transcription formation complex pathway [246] and known MS risk-associated susceptibility [247] interact with 25(OH)D to influence MS clinical

course. We also demonstrated that those genes involved in the vitamin D pathway interact with IFN-β influencing the levels of 25(OH)D in people with MS [248]. These results are presented in Chapter 4, Chapter 5 and Chapter 6.

As noted earlier, many studies have found a positive association between smoking and MS, including dose-response and duration of smoking. Meanwhile, the *HLA-A*02* allele may have a protective role for MS [236,249]. A Swedish study found that among those people with *HLA-DR15* and absence of *HLA-A*02* alleles, smokers were at significantly increased risk of MS (OR 13.5) compared to non-smokers with neither of these genetic risk factors [250]. The odds ratio for smokers without genetic risk was 1.4 and compared to 4.9 for non-smokers with both genetic risk factors. They also observed a significant interaction between absence of *HLA-A*02* and smoking among those carrying *HLA-DR15* alleles [250]. However, a pooled study including three MS case control studies showed no interaction between *HLA-DR15* and smoking [150], suggesting the need for further investigations to understand whether there is a true interaction and what potential mechanisms might underlie it.

Having younger siblings may be a marker of repeated exposure to common early life microbial infections. We found that low infant sibling exposure was associated with an increased risk of MS [251]. Moreover, we found an interaction between *HLA-DR15* and low infant sibling exposure [252]: the combined effect of *HLA-DR15* positivity and low infant sibling exposure on MS was nearly four-fold higher than expected, compared to the effects of *HLA-DR15* positivity and low infant sibling exposure alone.

Therefore there is now strong evidence to suggest that gene–environment interactions play a significant role in MS onset and may explain some of the missing heritability of

MS. Also, the investigations of gene–environment interactions have increased our understanding of how known environmental factors may influence gene expression and function and has linked known environmental and genetic factors together to provide biologically plausible mechanisms of how components of MS causation may occur.

1.7.5 Implications for MS patients and their families

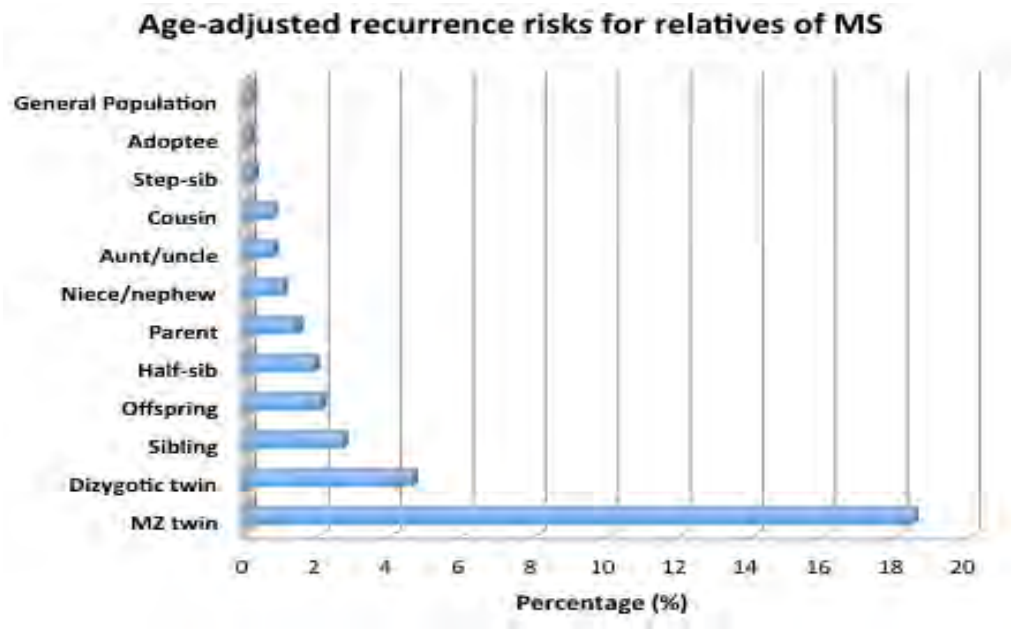
For people with MS, progressive disability substantially affects life expectancy and impairs quality of life. The median life expectancy for MS patients is five to ten years shorter than for age-matched general populations [253]. People with MS and their families are therefore very interested in anything that potentially may improve outcomes, may allow better and more accurate diagnosis of MS and may provide accurate information about the risk of MS among relatives, especially to children and siblings. Many people with MS, on being diagnosed, are alarmed to hear that there is a familial risk; they are keen to know whether relatives can be tested for MS risk and what can be done to reduce that risk. The revolution in our understanding of MS genetics that has occurred over the last five years; that is detailed above, can help to answer some of these questions. However there are several important points that clinicians need to be aware of when counseling people with MS.

- There is no genetic test that can absolutely define a person's risk of developing MS. The most highly associated genetic risk loci in the HLA region increases the risk of MS three-fold and is present in around 70% of people with MS and around 40% without MS. Therefore on an individual basis it provides little information.
- The other 63 described MS loci all contribute only a fraction of the risk and are all common alleles with the frequency seen in MS cases only slightly different from the normal population. Consequently they are not useful markers at the individual

level.

- The best test of increased risk is a family history of MS, with the risk increase for relatives now relatively well-understood (**Figure 1.6**). It is important to note that all relative risks are dependent on the background population risk. Therefore an accurate understanding of the population prevalence for the persons region of residence and ethnic group is essential. This can be estimated from the latitudinal data as described in the literature [75].
- The recent identification of the first MS rare variant in the *CYP27B1* gene may represent the first instance where knowledge of a genotype may influence outcomes. In multiplex families it is possible that testing for the described SNPs in the *CYP27B1* gene may be useful. However this has not been tested as yet and clearly more work is required before testing can be advocated.

Figure 1.6. Age-adjusted recurrent risk of MS for all relative group according to a recent meta-analysis [170]. The risk rank from highest to lowest: monozygotic twin (18.44%), dizygotic twin (4.61%), sibling (2.68 %), offspring (2.07%), half-sib (1.89%), parent (1.45%), niece/nephew (1.02%), aunt/uncle (0.75%), cousin (0.73%), step-sib (0.20%), adoptee (0.13%) and general population (0.16%).



Despite the dramatic increase in our knowledge of MS genetics, we have a long way to go. What is exciting is the rate at which we are learning. As noted in figure 1.3, until 2007 there was only one known MS genetic association and by 2012 we have information on over 60 loci. As genetic techniques become cheaper and more widely applicable, particularly whole-genome sequencing techniques, we will find more associations. These may well be rare variants with significant effects, such as the *CYP27B1* variants; they may be only present in a small fraction of people with MS and they may even be personal family mutations. However, each new causative locus is another chink in the armour of a previously impenetrable disease. Finding genetic loci that are associated opens a whole field of human-based research that allows the development of new treatments, allows us to understand how the genes and the environment interact and provides information that facilitates epigenetic approaches understanding the cause and progression of MS.

We cannot yet offer specific genetic testing to people with MS or their families. However, the finding of further rare variants may soon enable us to tailor treatments to families by

their genetic status. The genetic revolution has been a great step forward in MS research as it allows us to look at molecular mechanisms of this complicated disease in humans, rather than only in animals; this gives a far better chance of unraveling the cause of MS and developing preventative and treatment strategies.

1.7.6 Conclusion

Recent advances in genetic studies have confirmed that MS is a complex trait with genetic effects. Many environmental factors, and interactions between genes and environment, contribute to the development of MS. However, current knowledge of MS genetics is still insufficient and not yet clinically applicable. We cannot yet produce a predictive model of risk based on the genetic profile of an individual. However, gene identification efforts have considerably modified our understanding of the biological mechanisms underpinning the risk of and progression of MS. Promising approaches such as whole-exome and ultimately whole-genome sequencing have the potential to provide an exhaustive map of MS genes in the near future. We believe that the genetic information of MS can be meaningful, not only for scientists and clinicians, but also for MS patients and their families. In the long term, we are confident that progress in genetics research will lead to useful diagnostic and predictive tests and new treatments for MS.

1.8 Summary

MS is a chronic inflammatory degenerative disorder of the brain and spinal cord, leading to myelin and axonal damage caused by the individual’s inflammatory response. Three types of clinical course have been identified: relapsing remitting, secondary progressive and primary progressive. To date, there is no cure for the disease and in practice,

treatment varies with the stage reached in the course of the disease, but are mainly aimed at reducing the relapse rate and the accumulation of disability. The precise aetiology of MS is not known although research showed MS is driven by both environmental and genetic factors. So far, the best known and strongest environmental factors —lower ultraviolet radiation (UVR) exposure and/or lower vitamin D levels, Epstein Barr virus (EBV) infection and smoking have been well established. Additionally, 110 genetic variants have been identified conferring susceptibility to MS by GWAS, including the most significant HLA region and other about 100 non-HLA common variants. However, GWAS only explains a small fraction of the heritability of MS. Potential sources of the "missing heritability" include rare variants, many common variants with smaller effect, epigenetics, structure variants, gene-gene interactions, gene-environment interactions and pathway involvement. Using newer genetic technologies such as whole-exome sequencing, several rare functional variants have been identified as associated with MS. Promising approach such as whole-genome sequencing have the potential to provide an exhaustive map of MS genes in the near future. However, current knowledge of MS genetics is still insufficient and not yet clinically applicable. In the long term, we are confident that progress in genetics research will lead to useful diagnostic and predictive tests and new treatments for MS.

1.9 Aims of my PhD research

2. To detect rare genetic variants conferring susceptibility to MS using identity-by-decent mapping, so-called "population-based linkage analysis", utilising a large GWAS dataset comprising 3,543 cases and 5,898 controls. This will be presented in Chapter 2.

3. To assess the disease-causing variants to MS by family-based whole-genome sequencing analysis. This will be presented in Chapter 3.
4. To detect whether those genes involved in vitamin D metabolism and the VDR/RXR transcription factor formation complex pathway modify the relationship between 25-hydroxyvitamin D (25(OH)D) and relapse in people with MS, by using a well-validated prospective cohort study followed from 2002 to 2005. This will be presented in Chapter 4.
5. To detect whether known MS-associated susceptibility loci exert an effect on relapse and progression in clinical disability and whether these loci modify the association between 25(OH)D and relapse, by using a well-validated prospective cohort study followed from 2002 to 2005. This will be presented in Chapter 5.
6. To detect whether those genes involved in the vitamin D pathway modify the relationship between interferon-β and 25(OH)D levels in patients with MS by using a well-validated prospective cohort study followed from 2002 to 2005. This will be presented in Chapter 6.

Finally, Chapter 7 is presented a summary of main findings, an overarching discussion and future research.

1.10 Postscript

This chapter has provided some key information about MS, which will help to understand the rest of the thesis. The chapter has discussed the genetics of MS including rare variants. Rare functional variants are specifically important to the aetiology of MS, and different

methods have been developed to detect it. The next chapter will discuss how to detect rare variants by using identity-by-decent mapping with a large GWAS dataset.

1.11 References

1. (2008) World Health Organization: Atlas multiple sclerosis resources in the world 2008. Geneva: WHO Press.
2. Compston A, Coles A (2008) Multiple sclerosis. Lancet 372: 1502-1517.
3. Kurtzke JF (1983) Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33: 1444-1452.
4. Roxburgh RH, Seaman SR, Masterman T, Hensiek AE, Sawcer SJ, et al. (2005) Multiple Sclerosis Severity Score: using disability and disease duration to rate disease severity. Neurology 64: 1144-1151.
5. Sipe JC, Knobler RL, Braheny SL, Rice GP, Panitch HS, et al. (1984) A neurologic rating scale (NRS) for use in multiple sclerosis. Neurology 34: 1368-1372.
6. Lublin FD, Reingold SC (1996) Defining the clinical course of multiple sclerosis-results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 46: 907-911.
7. Rovaris M, Confavreux C, Furlan R, Kappos L, Comi G, et al. (2006) Secondary progressive multiple sclerosis: current knowledge and future challenges. Lancet Neurol 5: 343-354.
8. Reynolds R, Roncaroli F, Nicholas R, Radotra B, Gveric D, et al. (2011) The neuropathological basis of clinical progression in multiple sclerosis. Acta Neuropathol.
9. Lublin FD, Reingold SC (1996) Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 46: 907-911.
10. Tzartos JS, Friese MA, Craner MJ, Palace J, Newcombe J, et al. (2008) Interleukin-17 production in central nervous system-infiltrating T cells and glial cells is associated with active disease in multiple sclerosis. Am J Pathol 172: 146-155.
11. Kebir H, Kreymborg K, Ifergan I, Dodelet-Devillers A, Cayrol R, et al. (2007) Human TH17 lymphocytes promote blood-brain barrier disruption and central nervous system inflammation. Nat Med 13: 1173-1175.
12. Compston A, Coles A (2002) Multiple sclerosis. Lancet 359: 1221-1231.
13. Zajicek JP, Wing M, Scolding NJ, Compston DA (1992) Interactions between oligodendrocytes and microglia. A major role for complement and tumour necrosis factor in oligodendrocyte adherence and killing. Brain 115 (Pt 6): 1611-1631.
14. Kuhlmann T, Lingfeld G, Bitsch A, Schuchardt J, Bruck W (2002) Acute axonal damage in multiple sclerosis is most extensive in early disease stages and decreases over time. Brain 125: 2202-2212.

15. Kutzelnigg A, Lucchinetti CF, Stadelmann C, Bruck W, Rauschka H, et al. (2005) Cortical demyelination and diffuse white matter injury in multiple sclerosis. Brain 128: 2705-2712.
16. Patrikios P, Stadelmann C, Kutzelnigg A, Rauschka H, Schmidbauer M, et al. (2006) Remyelination is extensive in a subset of multiple sclerosis patients. Brain 129: 3165-3172.
17. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG (2000) Multiple sclerosis. N Engl J Med 343: 938–952.
18. Noronha A, Toscas A, Jensen MA (1993) Interferon beta decreases T cell activation and interferon gamma production in multiple sclerosis. J Neuroimmunol 46: 145-153.
19. Yong VW, Chabot S, Stuve O, Williams G (1998) Interferon beta in the treatment of multiple sclerosis: mechanisms of action. Neurology 51: 682-689.
20. Markowitz CE (2007) Interferon-beta: mechanism of action and dosing issues. Neurology 68: S8-11.
21. Jacobs LD, Cookfair DL, Rudick RA, Herndon RM, Richert JR, et al. (1996) Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group (MSCRG). Ann Neurol 39: 285-294.
22. PRISMS Study Group and the University of British Columbia MS/MRI Analysis Group. (2001) PRISMS-4: Long-term efficacy of interferon-beta-1a in relapsing MS. Neurology 56: 1628-1636.
23. Ebers CG, Group. PPoRaDbIβ-aSiMSS (1998) Randomised double-blind placebo-controlled study of interferon beta-1a in relapsing/remitting multiple sclerosis. PRISMS (Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis) Study Group. Lancet 352: 1498-1504.
24. Paty DW, Li DK (1993) Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. UBC MS/MRI Study Group and the IFNB Multiple Sclerosis Study Group. Neurology 43: 662-667.
25. The IFNB Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. (1995) Interferon beta-1b in the treatment of multiple sclerosis: final outcome of the randomized controlled trial. The IFNB Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. Neurology 45: 1277-1285.
26. Schmied M, Duda PW, Krieger JI, Trollmo C, Hafler DA (2003) In vitro evidence that subcutaneous administration of glatiramer acetate induces hyporesponsive T cells in patients with multiple sclerosis. Clin Immunol 106: 163-174.
27. Racke MK, Lovett-Racke AE (2011) Glatiramer acetate treatment of multiple sclerosis: an immunological perspective. J Immunol 186: 1887-1890.
28. Bornstein MB, Miller A, Slagle S, Weitzman M, Crystal H, et al. (1987) A pilot trial of Cop 1 in exacerbating-remitting multiple sclerosis. The New England journal of medicine 317: 408-414.
29. Johnson KP, Brooks BR, Cohen JA, Ford CC, Goldstein J, et al. (1995) Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: results of a phase III multicenter, double-blind placebo-controlled trial. The Copolymer 1 Multiple Sclerosis Study Group. Neurology 45: 1268-1276.
30. Ge Y, Grossman RI, Udupa JK, Fulton J, Constantinescu CS, et al. (2000) Glatiramer acetate (Copaxone) treatment in relapsing-remitting MS: quantitative MR assessment. Neurology 54: 813-817.

31. Comi G, Filippi M, Wolinsky JS (2001) European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging--measured disease activity and burden in patients with relapsing multiple sclerosis. European/Canadian Glatiramer Acetate Study Group. Annals of Neurology 49: 290-297.
32. Comi G, Martinelli V, Rodegher M, Moiola L, Bajenaru O, et al. (2009) Effect of glatiramer acetate on conversion to clinically definite multiple sclerosis in patients with clinically isolated syndrome (PreCISe study): a randomised, double-blind, placebo-controlled trial. Lancet 374: 1503-1511.
33. La Mantia L, Munari LM, Lovati R (2010) Glatiramer acetate for multiple sclerosis. Cochrane Database Syst Rev: CD004678.
34. Kinkel RP, Kollman C, O'Connor P, Murray TJ, Simon J, et al. (2006) IM interferon beta-1a delays definite multiple sclerosis 5 years after a first demyelinating event. Neurology 66: 678-684.
35. Kappos L, Freedman MS, Polman CH, Edan G, Hartung HP, et al. (2007) Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. Lancet 370: 389-397.
36. Panitch H, Miller A, Paty D, Weinshenker B (2004) Interferon beta-1b in secondary progressive MS: results from a 3-year controlled study. Neurology 63: 1788-1795.
37. Group. SPECToRI-B-aiMSS (2001) Randomized controlled trial of interferon- beta-1a in secondary progressive MS: Clinical results. Neurology 56: 1496-1504.
38. Cohen JA, Cutter GR, Fischer JS, Goodman AD, Heidenreich FR, et al. (2002) Benefit of interferon beta-1a on MSFC progression in secondary progressive MS. Neurology 59: 679-687.
39. Le Page E, Leray E, Taurin G, Coustans M, Chaperon J, et al. (2008) Mitoxantrone as induction treatment in aggressive relapsing remitting multiple sclerosis: treatment response factors in a 5 year follow-up observational study of 100 consecutive patients. J Neurol Neurosurg Psychiatry 79: 52-56.
40. Edan G, Miller D, Clanet M, Confavreux C, Lyon-Caen O, et al. (1997) Therapeutic effect of mitoxantrone combined with methylprednisolone in multiple sclerosis: a randomised multicentre study of active disease using MRI and clinical criteria. J Neurol Neurosurg Psychiatry 62: 112-118.
41. Hartung HP, Gonsette R, Konig N, Kwiecinski H, Guseo A, et al. (2002) Mitoxantrone in progressive multiple sclerosis: a placebo-controlled, double-blind, randomised, multicentre trial. Lancet 360: 2018-2025.
42. Wolinsky JS, Narayana PA, O'Connor P, Coyle PK, Ford C, et al. (2007) Glatiramer acetate in primary progressive multiple sclerosis: results of a multinational, multicenter, double-blind, placebo-controlled trial. Ann Neurol 61: 14-24.
43. Bochner BS, Luscinskas FW, Gimbrone MA, Jr., Newman W, Sterbinsky SA, et al. (1991) Adhesion of human basophils, eosinophils, and neutrophils to interleukin 1-activated human vascular endothelial cells: contributions of endothelial cell adhesion molecules. J Exp Med 173: 1553-1557.
44. Hemler ME, Elices MJ, Parker C, Takada Y (1990) Structure of the integrin VLA-4 and its cell-cell and cell-matrix adhesion functions. Immunol Rev 114: 45-65.
45. Lobb RR, Hemler ME (1994) The pathophysiologic role of alpha 4 integrins in vivo. J Clin Invest 94: 1722-1728.
46. Im DS (2003) Linking Chinese medicine and G-protein-coupled receptors. Trends Pharmacol Sci 24: 2-4.

47. Claussen MC, Korn T (2012) Immune mechanisms of new therapeutic strategies in MS: teriflunomide. Clin Immunol 142: 49-56.
48. Polman CH, O'Connor PW, Havrdova E, Hutchinson M, Kappos L, et al. (2006) A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. N Engl J Med 354: 899-910.
49. Cohen JA, Barkhof F, Comi G, Hartung HP, Khatri BO, et al. (2010) Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. N Engl J Med 362: 402-415.
50. Kappos L, Radue EW, O'Connor P, Polman C, Hohlfeld R, et al. (2010) A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. N Engl J Med 362: 387-401.
51. O'Connor P, Wolinsky JS, Confavreux C, Comi G, Kappos L, et al. (2011) Randomized trial of oral teriflunomide for relapsing multiple sclerosis. N Engl J Med 365: 1293-1303.
52. Giovannoni G (2011) Promising emerging therapies for multiple sclerosis. Neurol Clin 29: 435-448.
53. Kappos L, Gold R, Miller DH, MacManus DG, Havrdova E, et al. (2012) Effect of BG-12 on contrast-enhanced lesions in patients with relapsing--remitting multiple sclerosis: subgroup analyses from the phase 2b study. Mult Scler 18: 314-321.
54. Gold R, Kappos L, Arnold DL, Bar-Or A, Giovannoni G, et al. (2012) Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. N Engl J Med 367: 1098-1107.
55. Fox R (2012) Clinical Efficacy of BG-12 in Relapsing-Remitting Multiple Sclerosis (RRMS): Data from the Phase 3 CONFIRM Study. AAN New Orleans.
56. Albrecht P, Bouchachia I, Goebels N, Henke N, Hofstetter HH, et al. (2012) Effects of dimethyl fumarate on neuroprotection and immunomodulation. J Neuroinflammation 9: 163.
57. Coles AJ, Compston DA, Selmaj KW, Lake SL, Moran S, et al. (2008) Alemtuzumab vs. interferon beta-1a in early multiple sclerosis. N Engl J Med 359: 1786-1801.
58. Cohen JA, Coles AJ, Arnold DL, Confavreux C, Fox EJ, et al. (2012) Alemtuzumab versus interferon beta 1a as first-line treatment for patients with relapsing-remitting multiple sclerosis: a randomised controlled phase 3 trial. Lancet.
59. Coles AJ, Twyman CL, Arnold DL, Cohen JA, Confavreux C, et al. (2012) Alemtuzumab for patients with relapsing multiple sclerosis after disease-modifying therapy: a randomised controlled phase 3 trial. Lancet.
60. Bielekova B, Richert N, Howard T, Blevins G, Markovic-Plese S, et al. (2004) Humanized anti-CD25 (daclizumab) inhibits disease activity in multiple sclerosis patients failing to respond to interferon beta. Proc Natl Acad Sci U S A 101: 8705-8708.
61. Rose JW, Watt HE, White AT, Carlson NG (2004) Treatment of multiple sclerosis with an anti-interleukin-2 receptor monoclonal antibody. Ann Neurol 56: 864-867.
62. Rose JW, Burns JB, Bjorklund J, Klein J, Watt HE, et al. (2007) Daclizumab phase II trial in relapsing and remitting multiple sclerosis: MRI and clinical results. Neurology 69: 785-789.
63. Bielekova B, Howard T, Packer AN, Richert N, Blevins G, et al. (2009) Effect of anti-CD25 antibody daclizumab in the inhibition of inflammation and stabilization of disease progression in multiple sclerosis. Arch Neurol 66: 483-489.
64. Wynn D, Kaufman M, Montalban X, Vollmer T, Simon J, et al. (2010) Daclizumab in active relapsing multiple sclerosis (CHOICE study): a phase 2, randomised,

double-blind, placebo-controlled, add-on trial with interferon beta. Lancet Neurol 9: 381-390.
65. Kappos L, Li D, Calabresi PA, O'Connor P, Bar-Or A, et al. (2011) Ocrelizumab in relapsing-remitting multiple sclerosis: a phase 2, randomised, placebo-controlled, multicentre trial. Lancet 378: 1779-1787.
66. Hayes CE, Cantorna MT, DeLuca HF (1997) Vitamin D and multiple sclerosis. Proc Soc Exp Biol Med 216: 21-27.
67. Holick MF (2003) Vitamin D: A millenium perspective. J Cell Biochem 88: 296-307.
68. Holick MF (1994) McCollum Award Lecture, 1994: vitamin D--new horizons for the 21st century. Am J Clin Nutr 60: 619-630.
69. Zerwekh JE (2008) Blood biomarkers of vitamin D status. Am J Clin Nutr 87: 1087S-1091S.
70. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, et al. (2010) A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res 20: 1352-1360.
71. Hart PH, Gorman S, Finlay-Jones JJ (2011) Modulation of the immune system by UV radiation: more than just the effects of vitamin D? Nat Rev Immunol 11: 584-596.
72. Handel AE, Handunnetthi L, Giovannoni G, Ebers GC, Ramagopalan SV (2010) Genetic and environmental factors and the distribution of multiple sclerosis in Europe. Eur J Neurol 17: 1210-1214.
73. McLeod JG, Hammond SR, Hallpike JF (1994) Epidemiology of multiple sclerosis in Australia. With NSW and SA survey results. Med J Aust 160: 117-122.
74. Hernan MA, Olek MJ, Ascherio A (1999) Geographic variation of MS incidence in two prospective studies of US women. Neurology 53: 1711-1718.
75. Simpson S, Jr., Blizzard L, Otahal P, Van der Mei I, Taylor B (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. J Neurol Neurosurg Psychiatry 82: 1132-1141.
76. Alonso A, Hernan MA (2008) Temporal trends in the incidence of multiple sclerosis: a systematic review. Neurology 71: 129-135.
77. Ebers GC, Sadovnick AD (1993) The geographic distribution of multiple sclerosis: a review. Neuroepidemiology 12: 1-5.
78. Taylor BV, Pearson JF, Clarke G, Mason DF, Abernethy DA, et al. (2010) MS prevalence in New Zealand, an ethnically and latitudinally diverse country. Mult Scler 16: 1422-1431.
79. Freedman DM, Dosemeci M, Alavanja MC (2000) Mortality from multiple sclerosis and exposure to residential and occupational solar radiation: a case-control study based on death certificates. Occup Environ Med 57: 418-421.
80. Westberg M, Feychting M, Jonsson F, Nise G, Gustavsson P (2009) Occupational exposure to UV light and mortality from multiple sclerosis. Am J Ind Med 52: 353-357.
81. Grytten N, Torkildsen O, Aarseth JH, Benjaminsen E, Celius EG, et al. (2012) Month of birth as a latitude-dependent risk factor for multiple sclerosis in Norway. Mult Scler.
82. Willer CJ, Dyment DA, Sadovnick AD, Rothwell PM, Murray TJ, et al. (2005) Timing of birth and risk of multiple sclerosis: population based study. Bmj 330: 120.
83. Staples J, Ponsonby AL, Lim L (2010) Low maternal exposure to ultraviolet radiation in pregnancy, month of birth, and risk of multiple sclerosis in offspring: longitudinal analysis. Bmj 340: c1640.

84. Saastamoinen KP, Auvinen MK, Tienari PJ (2012) Month of birth is associated with multiple sclerosis but not with HLA-DR15 in Finland. Mult Scler 18: 563-568.
85. Bayes HK, Weir CJ, O'Leary C (2010) Timing of birth and risk of multiple sclerosis in the Scottish population. Eur Neurol 63: 36-40.
86. Duthie MS, Kimber I, Norval M (1999) The effects of ultraviolet radiation on the human immune system. Br J Dermatol 140: 995-1009.
87. Garssen J, van Loveren H (2001) Effects of ultraviolet exposure on the immune system. Crit Rev Immunol 21: 359-397.
88. Ascherio A, Munger KL, Simon KC (2010) Vitamin D and multiple sclerosis. Lancet Neurology 9: 599-612.
89. Lucas RM, Ponsonby AL (2006) Considering the potential benefits as well as adverse effects of sun exposure: can all the potential benefits be provided by oral vitamin D supplementation? Prog Biophys Mol Biol 92: 140-149.
90. Becklund BR, Severson KS, Vang SV, DeLuca HF (2010) UV radiation suppresses experimental autoimmune encephalomyelitis independent of vitamin D production. Proc Natl Acad Sci U S A 107: 6418-6423.
91. Lucas RM, Ponsonby AL, Dear K, Valery PC, Pender MP, et al. (2011) Sun exposure and vitamin D are independent risk factors for CNS demyelination. Neurology 76: 540-548.
92. van der Mei IA, Ponsonby AL, Dwyer T, Blizzard L, Taylor BV, et al. (2007) Vitamin D levels in people with multiple sclerosis and community controls in Tasmania, Australia. J Neurol 254: 581-590.
93. Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A (2006) Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. Jama 296: 2832-2838.
94. Kampman MT, Brustad M (2008) Vitamin D: a candidate for the environmental effect in multiple sclerosis - observations from Norway. Neuroepidemiology 30: 140-146.
95. Munger KL, Zhang SM, O'Reilly E, Hernan MA, Olek MJ, et al. (2004) Vitamin D intake and incidence of multiple sclerosis. Neurology 62: 60-65.
96. Bischoff-Ferrari H (2009) Vitamin D: what is an adequate vitamin D level and how much supplementation is necessary? Best Pract Res Clin Rheumatol 23: 789-795.
97. Looker AC, Pfeiffer CM, Lacher DA, Schleicher RL, Picciano MF, et al. (2008) Serum 25-hydroxyvitamin D status of the US population: 1988-1994 compared with 2000-2004. Am J Clin Nutr 88: 1519-1527.
98. Soilu-Hanninen M, Laaksonen M, Laitinen I, Eralinna JP, Lilius EM, et al. (2008) A longitudinal study of serum 25-hydroxyvitamin D and intact parathyroid hormone levels indicate the importance of vitamin D and calcium homeostasis regulation in multiple sclerosis. J Neurol Neurosurg Psychiatry 79: 152-157.
99. Soilu-Hanninen M, Airas L, Mononen I, Heikkila A, Viljanen M, et al. (2005) 25-Hydroxyvitamin D levels in serum at the onset of multiple sclerosis. Mult Scler 11: 266-271.
100. Correale J, Ysrraelit MC, Gaitan MI (2009) Immunomodulatory effects of Vitamin D in multiple sclerosis. Brain 132: 1146-1160.
101. Tremlett H, van der Mei IA, Pittas F, Blizzard L, Paley G, et al. (2008) Monthly ambient sunlight, infections and relapse rates in multiple sclerosis. Neuroepidemiology 31: 271-279.
102. Simpson S, Jr., Taylor B, Blizzard L, Ponsonby AL, Pittas F, et al. (2010) Higher 25-hydroxyvitamin D is associated with lower relapse risk in multiple sclerosis. Ann Neurol 68: 193-203.

103. Smolders J, Menheere P, Kessels A, Damoiseaux J, Hupperts R (2008) Association of vitamin D metabolite levels with relapse rate and disability in multiple sclerosis. Mult Scler 14: 1220-1224.
104. Runia TF, Hop WC, de Rijke YB, Buljevac D, Hintzen RQ (2012) Lower serum vitamin D levels are associated with a higher relapse risk in multiple sclerosis. Neurology 79: 261-266.
105. Mowry EM, Krupp LB, Milazzo M, Chabas D, Strober JB, et al. (2010) Vitamin D status is associated with relapse rate in pediatric-onset multiple sclerosis. Ann Neurol 67: 618-624.
106. Weinstock-Guttman B, Zivadinov R, Qu J, Cookfair D, Duan X, et al. (2011) Vitamin D metabolites are associated with clinical and MRI outcomes in multiple sclerosis patients. J Neurol Neurosurg Psychiatry 82: 189-195.
107. Neau JP, Artaud-Uriot MS, Lhomme V, Bounaud JY, Lebras F, et al. (2011) [Vitamin D and multiple sclerosis. A prospective survey of patients of Poitou-Charentes area]. Rev Neurol (Paris) 167: 317-323.
108. Mowry EM, Waubant E, McCulloch CE, Okuda DT, Evangelista AA, et al. (2012) Vitamin D status predicts new brain magnetic resonance imaging activity in multiple sclerosis. Ann Neurol 72: 234-240.
109. Holick MF, Chen TC (2008) Vitamin D deficiency: a worldwide problem with health consequences. Am J Clin Nutr 87: 1080S-1086S.
110. Prentice A (2008) Vitamin D deficiency: a global perspective. Nutr Rev 66: S153-164.
111. Hollis BW (2005) Circulating 25-hydroxyvitamin D levels indicative of vitamin D sufficiency: implications for establishing a new effective dietary intake recommendation for vitamin D. J Nutr 135: 317-322.
112. Dawson-Hughes B, Heaney RP, Holick MF, Lips P, Meunier PJ, et al. (2005) Estimates of optimal vitamin D status. Osteoporos Int 16: 713-716.
113. Vieth R (1999) Vitamin D supplementation, 25-hydroxyvitamin D concentrations, and safety. Am J Clin Nutr 69: 842-856.
114. Burton JM, Kimball S, Vieth R, Bar-Or A, Dosch HM, et al. (2010) A phase I/II dose-escalation trial of vitamin D3 and calcium in multiple sclerosis. Neurology 74: 1852-1859.
115. Stein MS, Liu Y, Gray OM, Baker JE, Kolbe SC, et al. (2011) A randomized trial of high-dose vitamin D2 in relapsing-remitting multiple sclerosis. Neurology 77: 1611-1618.
116. Soilu-Hanninen M, Aivo J, Lindstrom BM, Elovaara I, Sumelahti ML, et al. (2012) A randomised, double blind, placebo controlled trial with vitamin D3 as an add on treatment to interferon beta-1b in patients with multiple sclerosis. J Neurol Neurosurg Psychiatry 83: 565-571.
117. Ponsonby AL, van der Mei I, Dwyer T, Blizzard L, Taylor B, et al. (2005) Exposure to infant siblings during early life and risk of multiple sclerosis. Jama 293: 463-469.
118. Sumaya CV, Myers LW, Ellison GW (1980) Epstein-Barr virus antibodies in multiple sclerosis. Arch Neurol 37: 94-96.
119. Farrell RA, Antony D, Wall GR, Clark DA, Fisniku L, et al. (2009) Humoral immune response to EBV in multiple sclerosis is associated with disease activity on MRI. Neurology 73: 32-38.
120. DeLorenze GN, Munger KL, Lennette ET, Orentreich N, Vogelman JH, et al. (2006) Epstein-Barr virus and multiple sclerosis: evidence of association from a prospective study with long-term follow-up. Archives of Neurology 63: 839-844.

121. Zivadinov R, Nasuelli D, Tommasi MA, Serafin M, Bratina A, et al. (2006) Positivity of cytomegalovirus antibodies predicts a better clinical and radiological outcome in multiple sclerosis patients. Neurol Res 28: 262-269.
122. Shirodaria PV, Haire M, Fleming E, Merrett JD, Hawkins SA, et al. (1987) Viral antibody titers. Comparison in patients with multiple sclerosis and rheumatoid arthritis. Arch Neurol 44: 1237-1241.
123. Haahr S, Plesner AM, Vestergaard BF, Hollsberg P (2004) A role of late Epstein-Barr virus infection in multiple sclerosis. Acta Neurol Scand 109: 270-275.
124. Wandinger K, Jabs W, Siekhaus A, Bubel S, Trillenberg P, et al. (2000) Association between clinical disease activity and Epstein-Barr virus reactivation in MS. Neurology 55: 178-184.
125. Bray PF, Bloomer LC, Salmon VC, Bagley MH, Larsen PD (1983) Epstein-Barr virus infection and antibody synthesis in patients with multiple sclerosis. Arch Neurol 40: 406-408.
126. Larsen PD, Bloomer LC, Bray PF (1985) Epstein-Barr nuclear antigen and viral capsid antigen antibody titers in multiple sclerosis. Neurology 35: 435-438.
127. Sumaya CV, Myers LW, Ellison GW, Ench Y (1985) Increased prevalence and titer of Epstein-Barr virus antibodies in patients with multiple sclerosis. Ann Neurol 17: 371-377.
128. Riverol M, Sepulcre J, Fernandez-Diez B, Villoslada P, Fernandez-Alonso M, et al. (2007) Antibodies against Epstein-Barr virus and herpesvirus type 6 are associated with the early phases of multiple sclerosis. J Neuroimmunol 192: 184-185.
129. Khaki M, Ghazavi A, Ghasami K, Rafiei M, Payani MA, et al. (2011) Evaluation of viral antibodies in Iranian multiple sclerosis patients. Neurosciences (Riyadh) 16: 224-228.
130. Ascherio A, Munger KL (2007) Environmental risk factors for multiple sclerosis. Part I: the role of infection. Ann Neurol 61: 288-299.
131. Ascherio A, Munch M (2000) Epstein-Barr virus and multiple sclerosis. Epidemiology 11: 220-224.
132. Ascherio A, Munger KL, Lennette ET, Spiegelman D, Hernan MA, et al. (2001) Epstein-Barr virus antibodies and risk of multiple sclerosis: a prospective study. Jama 286: 3083-3088.
133. Sundstrom P, Juto P, Wadell G, Hallmans G, Svenningsson A, et al. (2004) An altered immune response to Epstein-Barr virus in multiple sclerosis: a prospective study. Neurology 62: 2277-2282.
134. Levin LI, Munger KL, Rubertone MV, Peck CA, Lennette ET, et al. (2005) Temporal relationship between elevation of epstein-barr virus antibody titers and initial onset of neurological symptoms in multiple sclerosis. Jama 293: 2496-2500.
135. Kusunoki Y, Huang H, Fukuda Y, Ozaki K, Saito M, et al. (1993) A positive correlation between the precursor frequency of cytotoxic lymphocytes to autologous Epstein-Barr virus-transformed B cells and antibody titer level against Epstein-Barr virus-associated nuclear antigen in healthy seropositive individuals. Microbiol Immunol 37: 461-469.
136. Henle W, Henle G, Niederman JC, Klemola E, Haltia K (1971) Antibodies to early antigens induced by Epstein-Barr virus in infectious mononucleosis. J Infect Dis 124: 58-67.
137. Horwitz CA, Henle W, Henle G, Rudnick H, Latts E (1985) Long-term serological follow-up of patients for Epstein-Barr virus after recovery from infectious mononucleosis. J Infect Dis 151: 1150-1153.

138. Levin LI, Munger KL, Rubertone MV, Peck CA, Lennette ET, et al. (2003) Multiple sclerosis and Epstein-Barr virus. Jama 289: 1533-1536.
139. Thacker EL, Mirzaei F, Ascherio A (2006) Infectious mononucleosis and risk for multiple sclerosis: a meta-analysis. Ann Neurol 59: 499-503.
140. Handel AE, Williamson AJ, Disanto G, Handunnetthi L, Giovannoni G, et al. (2010) An updated meta-analysis of risk of multiple sclerosis following infectious mononucleosis. PLoS One 5.
141. Holmoy T, Vartdal F (2004) Cerebrospinal fluid T cells from multiple sclerosis patients recognize autologous Epstein-Barr virus-transformed B cells. J Neurovirol 10: 52-56.
142. Lunemann JD, Edwards N, Muraro PA, Hayashi S, Cohen JI, et al. (2006) Increased frequency and broadened specificity of latent EBV nuclear antigen-1-specific T cells in multiple sclerosis. Brain 129: 1493-1506.
143. Serafini B, Rosicarelli B, Franciotta D, Magliozzi R, Reynolds R, et al. (2007) Dysregulated Epstein-Barr virus infection in the multiple sclerosis brain. J Exp Med 204: 2899-2912.
144. Tzartos JS, Khan G, Vossenkamper A, Cruz-Sadaba M, Lonardi S, et al. (2012) Association of innate immune activation with latent Epstein-Barr virus in active MS lesions. Neurology 78: 15-23.
145. Willis SN, Stadelmann C, Rodig SJ, Caron T, Gattenloehner S, et al. (2009) Epstein-Barr virus infection is not a characteristic feature of multiple sclerosis brain. Brain 132: 3318-3328.
146. Sargsyan SA, Shearer AJ, Ritchie AM, Burgoon MP, Anderson S, et al. (2010) Absence of Epstein-Barr virus in the brain and CSF of patients with multiple sclerosis. Neurology 74: 1127-1135.
147. Tselis A (2012) Epstein-Barr virus cause of multiple sclerosis. Curr Opin Rheumatol 24: 424-428.
148. Marrie RA (2004) Environmental risk factors in multiple sclerosis aetiology. Lancet Neurol 3: 709-718.
149. Simpson CA, Newell DJ, Schapira K (1966) Smoking and multiple sclerosis. Neurology 16: 1041-1043 passim.
150. Simon KC, van der Mei IA, Munger KL, Ponsonby A, Dickinson J, et al. (2010) Combined effects of smoking, anti-EBNA antibodies, and HLA-DRB1*1501 on multiple sclerosis risk. Neurology 74: 1365-1371.
151. Hernan MA, Olek MJ, Ascherio A (2001) Cigarette smoking and incidence of multiple sclerosis. Am J Epidemiol 154: 69-74.
152. Carlens C, Hergens MP, Grunewald J, Ekbom A, Eklund A, et al. (2010) Smoking, use of moist snuff, and risk of chronic inflammatory diseases. Am J Respir Crit Care Med 181: 1217-1222.
153. Hernan MA, Jick SS, Logroscino G, Olek MJ, Ascherio A, et al. (2005) Cigarette smoking and the progression of multiple sclerosis. Brain 128: 1461-1465.
154. Healy BC, Ali EN, Guttmann CR, Chitnis T, Glanz BI, et al. (2009) Smoking and disease progression in multiple sclerosis. Arch Neurol 66: 858-864.
155. Sundstrom P, Nystrom L (2008) Smoking worsens the prognosis in multiple sclerosis. Mult Scler 14: 1031-1035.
156. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11: 2417-2423.
157. Pittas F, Ponsonby AL, van der Mei IA, Taylor BV, Blizzard L, et al. (2009) Smoking is associated with progressive disease course and increased progression

in clinical disability in a prospective cohort of people with multiple sclerosis. J Neurol 256: 577-585.
158. Sopori M (2002) Effects of cigarette smoke on the immune system. Nat Rev Immunol 2: 372-377.
159. Sopori ML, Kozak W (1998) Immunomodulatory effects of cigarette smoke. J Neuroimmunol 83: 148-156.
160. Hedstrom AK, Baarnhielm M, Olsson T, Alfredsson L (2009) Tobacco smoking, but not Swedish snuff use, increases the risk of multiple sclerosis. Neurology 73: 696-701.
161. Fusby JS, Kassmeier MD, Palmer VL, Perry GA, Anderson DK, et al. (2010) Cigarette smoke-induced effects on bone marrow B-cell subsets and CD4+:CD8+ T-cell ratios are reversed by smoking cessation: influence of bone mass on immune cell response to and recovery from smoke exposure. Inhal Toxicol 22: 785-796.
162. Motz GT, Eppert BL, Wesselkamper SC, Flury JL, Borchers MT (2010) Chronic cigarette smoke exposure generates pathogenic T cells capable of driving COPD-like disease in Rag2-/- mice. Am J Respir Crit Care Med 181: 1223-1233.
163. Motz GT, Eppert BL, Wortham BW, Amos-Kroohs RM, Flury JL, et al. (2010) Chronic cigarette smoke exposure primes NK cell activation in a mouse model of chronic obstructive pulmonary disease. J Immunol 184: 4460-4469.
164. Kallberg H, Ding B, Padyukov L, Bengtsson C, Ronnelid J, et al. (2011) Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. Ann Rheum Dis 70: 508-511.
165. Harel-Meir M, Sherer Y, Shoenfeld Y (2007) Tobacco smoking and autoimmune rheumatic diseases. Nat Clin Pract Rheumatol 3: 707-715.
166. Wingerchuk DM (2012) Smoking: effects on multiple sclerosis susceptibility and disease progression. Ther Adv Neurol Disord 5: 13-22.
167. Handel AE, Williamson AJ, Disanto G, Dobson R, Giovannoni G, et al. (2011) Smoking and multiple sclerosis: an updated meta-analysis. PLoS One 6: e16149.
168. Rosati G (2001) The prevalence of multiple sclerosis in the world: an update. Neurol Sci 22: 117-139.
169. Robertson NP, Fraser M, Deans J, Clayton D, Walker N, et al. (1996) Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. Brain 119 (Pt 2): 449-455.
170. O'Gorman C, Lin R, Stankovich J, Broadley SA (2012) Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data. Neuroepidemiology 40: 1-12.
171. Sadovnick AD, Armstrong H, Rice GP, Bulman D, Hashimoto L, et al. (1993) A population-based study of multiple sclerosis in twins: update. Ann Neurol 33: 281-285.
172. Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, et al. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 11: 504-510.
173. Ebers GC, Sadovnick AD, Risch NJ (1995) A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. Nature 377: 150-151.
174. Sadovnick AD, Ebers GC, Dyment DA, Risch NJ (1996) Evidence for genetic basis of multiple sclerosis. The Canadian Collaborative Study Group. Lancet 347: 1728-1730.

175. Robertson NP, O'Riordan JI, Chataway J, Kingsley DP, Miller DH, et al. (1997) Offspring recurrence rates and clinical characteristics of conjugal multiple sclerosis. Lancet 349: 1587-1590.
176. Jersild C, Dupont B, Fog T, Platz PJ, Svejgaard A (1975) Histocompatibility determinants in multiple sclerosis. Transplant Rev 22: 148-163.
177. Jersild C, Svejgaard A, Fog T (1972) HL-A antigens and multiple sclerosis. Lancet 1: 1240-1241.
178. Eraksoy M, Kurtuncu M, Akman-Demir G, Kilinc M, Gedizlioglu M, et al. (2003) A whole genome screen for linkage in Turkish multiple sclerosis. J Neuroimmunol 143: 17-24.
179. Hensiek AE, Roxburgh R, Smilie B, Coraddu F, Akesson E, et al. (2003) Updated results of the United Kingdom linkage-based genome screen in multiple sclerosis. J Neuroimmunol 143: 25-30.
180. Kenealy SJ, Babron MC, Bradford Y, Schnetz-Boutaud N, Haines JL, et al. (2004) A second-generation genomic screen for multiple sclerosis. Am J Hum Genet 75: 1070-1078.
181. Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, et al. (2005) A high-density screen for linkage in multiple sclerosis. Am J Hum Genet 77: 454-467.
182. International HapMap Consortium. (2003) The International HapMap Project. Nature 426: 789-796.
183. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-2229.
184. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
185. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
186. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-913.
187. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet.
188. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
189. Ban M, McCauley JL, Zuvich R, Baker A, Bergamaschi L, et al. (2010) A non-synonymous SNP within membrane metalloendopeptidase-like 1 (MMEL1) is associated with multiple sclerosis. Genes Immun 11: 660-664.
190. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851-862.
191. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
192. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41: 776-782.

193. International Multiple Sclerosis Genetics Consortium (IMSGC). (2010) Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 19: 953-962.
194. Patsopoulos NA, de Bakker PI, Esposito F, Reischl J, Lehr S, et al. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol 70: 897-912.
195. Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, et al. (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 42: 495-497.
196. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767-778.
197. Jakkula E, Leppa V, Sulonen AM, Varilo T, Kallio S, et al. (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am J Hum Genet 86: 285-291.
198. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2009) CD226 Gly307Ser association with multiple autoimmune diseases. Genes Immun 10: 5-10.
199. Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, et al. (2009) Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet 17: 1309-1313.
200. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847-856.
201. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118-1125.
202. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.
203. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
204. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.
205. Ramagopalan SV, Dyment DA, Cader MZ, Morrison KM, Disanto G, et al. (2011) Rare variants in the CYP27B1 gene are associated with multiple sclerosis. Ann Neurol 70: 881-886.
206. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI (2011) Evolutionary evidence of the effect of rare variants on disease etiology. Clin Genet 79: 199-206.
207. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695-701.
208. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387-389.
209. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72: 1117-1130.

210. Davidson S, Starkey A, MacKenzie A (2009) Evidence of uneven selective pressure on different subsets of the conserved human genome; implications for the significance of intronic and intergenic DNA. BMC Genomics 10: 614.
211. Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK (2007) Recent acceleration of human adaptive evolution. Proc Natl Acad Sci U S A 104: 20753-20758.
212. Liu J, Zhang Y, Lei X, Zhang Z (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol 9: R69.
213. Eyre-Walker A (2010) Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A 107 Suppl 1: 1752-1756.
214. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.
215. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591-597.
216. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100-112.
217. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
218. Murphy R, Ellard S, Hattersley AT (2008) Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. Nat Clin Pract Endocrinol Metab 4: 200-213.
219. Hornum L, Romer J, Markholst H (2002) The diabetes-prone BB rat carries a frameshift mutation in Ian4, a positional candidate of Iddm1. Diabetes 51: 1972-1979.
220. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106: 19096-19101.
221. Browning SR (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178: 2123-2132.
222. Dyment DA, Cader MZ, Chao MJ, Lincoln MR, Morrison KM, et al. (2012) Exome sequencing identifies a novel multiple sclerosis susceptibility variant in the TYK2 gene. Neurology 79: 406-411.
223. Couturier N, Bucciarelli F, Nurtdinov RN, Debouverie M, Lebrun-Frenay C, et al. (2011) Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. Brain 134: 693-703.
224. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33 Suppl: 245-254.
225. Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. Science 187: 226-232.
226. Probst AV, Dunleavy E, Almouzni G (2009) Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol 10: 192-206.
227. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057-1068.

228. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, et al. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A 102: 10604-10609.
229. Szyf M, McGowan P, Meaney MJ (2008) The social environment and the epigenome. Environ Mol Mutagen 49: 46-60.
230. Hoppenbrouwers IA, Liu F, Aulchenko YS, Ebers GC, Oostra BA, et al. (2008) Maternal transmission of multiple sclerosis in a dutch population. Arch Neurol 65: 345-348.
231. Ramagopalan SV, Yee IM, Dyment DA, Orton SM, Marrie RA, et al. (2009) Parent-of-origin effect in multiple sclerosis: observations from interracial matings. Neurology 73: 602-605.
232. Chao MJ, Herrera BM, Ramagopalan SV, Deluca G, Handunetthi L, et al. (2010) Parent-of-origin effects at the major histocompatibility complex in multiple sclerosis. Hum Mol Genet 19: 3679-3689.
233. Kantarci OH, Barcellos LF, Atkinson EJ, Ramsay PP, Lincoln R, et al. (2006) Men transmit MS more often to their children vs women: the Carter effect. Neurology 67: 305-310.
234. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, et al. (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464: 1351-1356.
235. Handunnetthi L HA, Ramagopalan SV. (2010) Contribution of genetic, epigenetic and transcriptomic differences to twin discordance in multiple sclerosis. Expert Rev Neurother 10: 1379-1381.
236. Brynedal B, Duvefelt K, Jonasdottir G, Roos IM, Akesson E, et al. (2007) HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. PLoS One 2: e664.
237. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18: 2078-2090.
238. Lunemann JD, Kamradt T, Martin R, Munz C (2007) Epstein-barr virus: environmental trigger of multiple sclerosis? J Virol 81: 6777-6784.
239. Kruger S, Schroers R, Rooney CM, Gahn B, Chen SY (2003) Identification of a naturally processed HLA-DR-restricted T-helper epitope in Epstein-Barr virus nuclear antigen type 1. J Immunother 26: 212-221.
240. van der Mei IA, Simpson S, Jr., Stankovich J, Taylor BV (2011) Individual and joint action of environmental factors and risk of MS. Neurol Clin 29: 233-255.
241. Tajouri L, Ovcaric M, Curtain R, Johnson MP, Griffiths LR, et al. (2005) Variation in the vitamin D receptor gene is associated with multiple sclerosis in an Australian population. J Neurogenet 19: 25-38.
242. Dickinson JL, Perera DI, van der Mei AF, Ponsonby AL, Polanowski AM, et al. (2009) Past environmental sun exposure and risk of multiple sclerosis: a role for the Cdx-2 Vitamin D receptor variant in this interaction. Mult Scler 15: 563-570.
243. Simon KC, Munger KL, Xing Y, Ascherio A (2010) Polymorphisms in vitamin D metabolism related genes and risk of multiple sclerosis. Mult Scler 16: 133-138.
244. Ramagopalan SV, Maugeri NJ, Handunnetthi L, Lincoln MR, Orton SM, et al. (2009) Expression of the multiple sclerosis-associated MHC class II Allele HLA-DRB1*1501 is regulated by vitamin D. PLoS Genet 5: e1000369.
245. Ramagopalan SV, Link J, Byrnes JK, Dyment DA, Giovannoni G, et al. (2009) HLA-DRB1 and month of birth in multiple sclerosis. Neurology 73: 2107-2111.

246. Lin R, Taylor BV, Simpson S, Jr., Charlesworth J, Ponsonby AL, et al. (2013) Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis. J Neurol Neurosurg Psychiatry.
247. Lin R, Taylor BV, Simpson S, Jr., Charlesworth J, Ponsonby AL, et al. (2013) Association between multiple sclerosis risk-associated SNPs and relapse and disability - a prospective cohort study. Mult Scler.
248. Lin R, Taylor BV, Charlesworth J, van der Mei IA, Blizzard L, et al. (2013) Modulating effects of the gene WT1 on the interferon-β-vitamin D association in MS. Multiple sclerosis, submitted.
249. Bergamaschi L, Leone MA, Fasano ME, Guerini FR, Ferrante D, et al. (2010) HLA-class I markers and multiple sclerosis susceptibility in the Italian population. Genes Immun 11: 173-180.
250. Hedstrom AK, Sundqvist E, Baarnhielm M, Nordin N, Hillert J, et al. (2011) Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. Brain 134: 653-664.
251. Ponsonby AL, van der Mei I, Dwyer T, Blizzard L, Taylor B, et al. (2005) Exposure to infant siblings during early life and risk of multiple sclerosis. Jama 293: 463-469.
252. van der Mei IA, Ponsonby AL, Taylor BV, Stankovich J, Dickinson JL, et al. (2010) Human leukocyte antigen-DR15, low infant sibling exposure and multiple sclerosis: gene-environment interaction. Ann Neurol 67: 261-265.
253. Bronnum-Hansen H, Koch-Henriksen N, Stenager E (2004) Trends in survival and cause of death in Danish patients with multiple sclerosis. Brain 127: 844-850.

The following article has been removed for copyright or proprietary reasons

Appendix 1A: Publication in Chapter 1

Rui Lin, Jac Charlesworth, Ingrid van der Mei, Bruce V Taylor. The genetics of multiple sclerosis. *Practical Neurology.* 2012;12(5):279-288. (Invited review)

Chapter 2. Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis

2.1 Preface

The previous chapter provided a general background of MS, including the environmental and genetic factors associated with the onset of the disease. Among the genetic factors, we particularly mentioned that there was 'missing heritability" that is heritable risk factors not yet been identified, including rare variants. This chapter will particularly focus on detecting genetic rare variants which are recognised to potentially have larger effects on complex disease compared to genetic common variants. Most parts of this chapter have been published in PLOS ONE 2013; 8(3): e56379 (Appendix 2A), whereas text marked as grey boxes are added and were not part of the original publication.

2.2 Introduction

Multiple sclerosis is a complex neurological disease of the central nervous system (CNS) triggered by environmental and genetic factors. There is considerable evidence for a significant genetic component to MS susceptibility, such as a higher concordance rate in monozygotic twins (24%-30%) than dizygotic twins (3%-5%) [1,2]. As for other immune diseases, genome-wide association studies (GWASs) have been highly successful for MS: uncovering about 110 common genetic variants associated with disease [3-14]. The majority of these variants lie near genes with known functions in the immune system and

these variants have also been associated with other autoimmune diseases, often in the opposite direction [15]. Virtually all of the variants confer modest increases in disease risk, the outstanding exception being the strong association with the *HLA-DRB1*15:01* allele in the major histocompatibility complex (MHC), which was first detected in the 1970's [16,17].

Despite this success, the variants identified by GWASs to date only explain 18-24% of the heritability of MS [13,18]. While much of the missing heritability is probably explained by common variants of even smaller effect sizes, some heritability may be explained by rare variants of larger effect size. Standard analysis of GWAS data is not designed to detect associations with rare variants that many believe may play an important role in the aetiology of complex traits [19-21]. Interestingly, GWASs have had less success for putative neurodegenerative diseases, such as Parkinson's disease, than for MS. For these diseases, family-based approaches detecting rare variants have been more successful [22,23]. This raises the possibility that rare variants under negative selection pressure are relatively more important in the genetic architecture of neurodegenerative processes, whereas common variants under balancing selection are more important in the genetic architecture of immunological processes. Discovery of rare MS susceptibility variants may alter perspectives on the relative importance of immunological & neurodegenerative processes in MS onset.

Standard analyses of GWAS data are not designed to detect associations with low frequency variants (MAF≤ 5%), and other strategies are required. One approach is to re-sequence loci containing common susceptibility variants identified from GWASs. This strategy was used to detect rare variants in *IFIH1* conferring protection to type I diabetes

[19]. However this strategy precludes the identification of new loci. Eventually it will be possible to overcome this limitation by whole genome sequencing, but it remains prohibitively expensive to perform adequately-powered studies. An alternative is to re-analyse GWAS data using identity-by-descent (IBD) mapping [24], also referred to as 'population-based linkage analysis' (PBLA) [25]. PBLA describes linkage analysis applied at the population level to detect mega base-scale regions where cases have inherited long haplotypes from distant ancestors, 10–100 generations ago. IBD mapping is performed on the unrelated individuals to determine whether these mega-base scale regions are identical and inherited from a common ancestor. If the common ancestor lived more than ten generations ago the individuals will share very short tracts of genetic material, and a shared haplotype that is very rare is also very likely to be IBD. HapMap Phase 3 identified that lower frequency variants should, on average, be younger than more common variants; and thus display a greater extent of haplotype sharing [26]. Therefore, if case pairs can be detected with long shared haplotypes (generally one to five megabases) inherited from distant common ancestors, then rare variants influencing disease risk can be localised. Even when whole genome sequencing becomes cheap enough to pursue with substantial sample sizes, IBD mapping may still help reduce the massive multiple testing problem by prioritizing regions. This is similar to the technique of prioritising association signals in regions of linkage [27].

Several methods of IBD mapping have been published: these include PLINK [25], GERMLINE [28], BEAGLE IBD [29] and BEAGLE fastIBD [30]. The models employed by PLINK and GERMLINE assume SNPs are in linkage equilibrium, and so 'pruning' of SNPs [25] is required to avoid false positives due to under-estimates of population haplotype frequencies. However pruning of SNPs in incomplete linkage disequilibrium

(LD) discards potentially useful information and reduces power. BEAGLE IBD and fastIBD implement a variable length Hidden Markov Model [31] to account for LD and model haplotype frequencies more accurately. BEAGLE fastIBD runs considerably faster than BEAGLE IBD (of the order of 1000 times faster with large GWAS datasets). This is mainly because 1) it does not formally model IBD status ('IBD'/'not IBD') between pairs of individuals using a Hidden Markov Model as in BEAGLE IBD; 2) it stores haplotype frequencies in a data dictionary (as in GERMLINE) which means computational time scales with sample size n like n log n instead of n^2.

To detect MS rare variants, we here use BEAGLE fastIBD to perform an IBD analysis on several large MS GWAS datasets comprised 3543 cases and 5898 controls. We identified a region of high significance on chromosome 19q13.43, with a genome-wide significant localisation signal ($p= 1.9\times10^{-6}$; LOD=4.65) using thresholds based on IBD segment length greater than 3cM and the probability p-value less than 10^{-9} (3cM_1e-9). This locus was deemed genome-wide significant according to the recently established genome-wide significance thresholds set for IBD mapping [32]. Analysis of expression data and investigation of genes in this area support the hypothesis for regulation of gene expression in this region to impact upon development or health of CNS tissue. Our analyses also illustrate some of the practical issues to deal with in IBD analyses, and demonstrate that IBD mapping can form a potentially powerful method for detecting rare variants in unrelated individuals at the population level.

2.3 Methods

2.3.1 Study subjects

All the MS cases and controls were recruited and genotyped from MS GWASs totaling 3,543 cases and 5,898 controls. Of these, 1,618 cases and 3,413 controls were from an Australian and New Zealand MS GWAS conducted by the Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) [3], and those DNA samples were genotyped on the Illumina Infinium Hap370CNV array [3]. An additional 861 Australian and New Zealand MS cases were genotyped with the Illumina Human660-Quad chip as part of a GWAS performed by the International MS Genetics Consortium (IMSGC) and the Wellcome Trust Case Control Consortium-2 (WTCCC2) [13]. Controls included 1,531 unrelated Australian samples from a GWAS genotyped by Queensland Institute of Medical Research (QIMR) with the Illumina Human610-Quad chip [33], and 1064 MS cases and 954 controls genotyped with the Sentrix® HumanHap550 BeadChip from a GWAS conducted in the US (GeneMSA) [6] [accessed via dbGAP].

2.3.2 Quality control of data

Conservative quality control measures were imposed both on the individual datasets before merging, and in the combined dataset after merging: SNPs with call rates less than 0.95 or in Hardy-Weinberg disequilibrium ($p<10^{-7}$) were discarded, as were samples with call rates less than 0.98. Duplicates and close relatives were also removed. This data cleaning was performed using PLINK [25].

A principal components analysis (PCA) was conducted by EIGENSTRAT [34] to exclude ancestry outliers and examine population structure within the remaining samples. First, SNPs in strong LD were pruned (using the PLINK --indep command with options 50 5

1.5), and then we excluded previously identified regions of high LD [35]. Outliers in the PCA were excluded using standard settings in Eigenstrat (more than six standard deviations from the mean along the first 10 principal components). All chromosomal locations refer to Human genome version hg18.

2.3.3 Running BEAGLE fastIBD and results processing

The fastIBD analyses were conducted using BEAGLE (http://faculty.washington.edu/browning/beagle/). In brief, genotypes for the merged, cleaned dataset were converted to BEAGLE format by using the linkage2Beagle.jar utility program. We then used the BEAGLE method for phasing the data and identifying IBD segments simultaneously, using the 'fastibdthreshold' option. This procedure was run 10 times for each chromosome starting with different seeds of the random number generator.

The output of these calculations was a series of "putative" IBD segments shared between pairs of individuals. Each segment comes with the following information attached: ids for the pair of individuals, first and last SNPs in the IBD segment, length of the segment in centimorgans, and probability of the two individuals both carrying the segment if it was not IBD. We filtered these segments using various maximum probabilities and minimum segment lengths, as recommended in the BEAGLE manual. Results from the 10 runs were combined by taking the union of IBD segments detected in each run. From the final list of segments, we wrote a Perl script to count numbers of case-case pairs (*yi*), case-control pairs (*ui*) and control-control pairs (*vi*) estimated to share haplotypes IBD at each SNP *i.*

2.3.4 Analysis of IBD

We focused on the detection of loci where groups of cases have inherited rare susceptibility alleles IBD. To do this, we modelled IBD sharing *yi* in case-case pairs ("case pairs") as a function of IBD sharing in $xi = ui + vi$ in case-control pairs and control-control pairs combined ("control pairs").

We tried various methods to model the *yi* as a function of the *xi*: linear regression, negative binomial regression and Poisson regression. Models were fitted using R [36], and goodness of fit was assessed by examining diagnostic plots [**Supplementary R commands**].

At SNPs *i* with more IBD sharing in cases than expected, residuals *zi* from the fitted models should be large and positive. To present residuals on a scale more familiar to geneticists, we converted them to LOD scores using the formula $LODi = zi^2/(2*\log_e(10))$.

At the SNPs with the highest LOD scores, we calculated the proportions of case pairs sharing IBD in various populations, and plotted networks of case and control pairs sharing IBD with each other using the R network package (http://cran.rproject.org/web/packages/network/index.html).

2.4 Results

2.4.1 Study samples from GWASs after cleaning

11 individuals were excluded because of call rates less than 0.98 and 202 were excluded because they were close relatives or duplicates. PCA was conducted on a subset of 77,856 SNPs not in LD, which were common to all sample sets. Through successive iterations 251 outliers (37 AUS cases; 9 AUS controls; 6 UK controls; 97 US cases; 102 US controls) were excluded. All datasets overlap well after removing outliers (**Figure 2.1**). In summary, there were 3,243 cases and 5,725 controls with 274,735 autosomal SNPs in the final analysis (**Table 2.1)**.

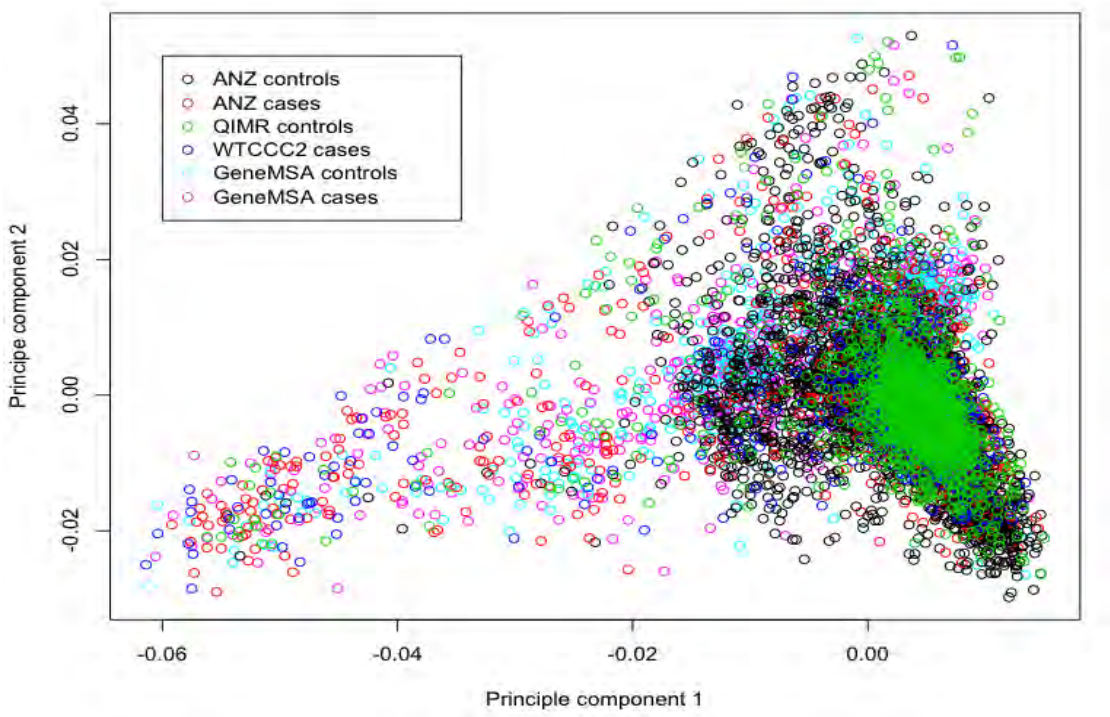
Table 2. 1. Sample numbers from GWASs (after cleaning)

GWAS Dataset	Country of origin		No. Case	No. Control	Total	No. SNPs
	Cases	Controls				
ANZgene[3]	AUS, NZ	US, UK	1,608	3,404	5,012	300,900
WTCCC2[13]	AUS	-	766	-	766	586,393
QIMR[33]	-	AUS	-	1,516	1,516	529,292
GeneMSA [6]	US	US	878	805	1,683	550,677
Total			3,252	5,725	8,977	274,735*

* The number of SNPs that passed QC in all 4 GWAS datasets.

AUS = Australia, NZ=New Zealand

Figure 2.1 Principal components analysis of the dataset. Most individuals in the dataset are of predominantly northern European ancestry (right hand side), but some have southern European ancestry (left hand side) (one dot for each individual).



2.4.2 Results of IBD analysis

We detected IBD with the threshold of IBD segment greater than 3cM and the haplotype probability p-value less than 10^{-9} (3cM_1e-9). A strong linkage signal was observed in the HLA region (LOD = 3.58), while the strongest signal in non-HLA region was on chromosome 19 (LOD =4.65), which reached genome-wide significance according to the recent established genome-wide significance threshold set for IBD mapping [32].

Results note 2.1 Example of IBD processing results

The detail of IBD process was not provided in our original published paper. To well understand the whole process, we added the example here. **Note 2.1. Table A** presents the results of IBD detected on chromosome 1 with the threshold of 3cM_1e-9 as an example. This table presents the begin marker (SNP) and the end mark (SNP) for each case_case pairs (*yi*), case_control pairs (*ui*) and control_control pairs (*vi*) in each run. Then we united the IBD segments in the ten runs because the IBD segments detected in each run would overlap each other. Finally we noted how many case_case (*yi*), case_control (*ui*) and control_control (*vi*) pairs shared IBD segments in each SNP in whole genome (**Note 2.1. Table B**).

Results note 2.1. Table A. Results of IBD detected on chromosome 1 with the threshold of 3cM_1e-9

ID1	ID2	Case/control (code)*	Case/control (code)*	Chr	Begin marker	End marker
Case1	Case2	2	2	1	4802	5212
Case1	Case3	2	2	1	4719	5212
Case1	Control1	2	1	1	2573	2976
Case1	Control2	2	1	1	2916	3132
Control2	Conrol3	1	1	1	19714	19900
Control2	Control4	1	1	1	19715	19905
…						

* 2: case; 1: control.

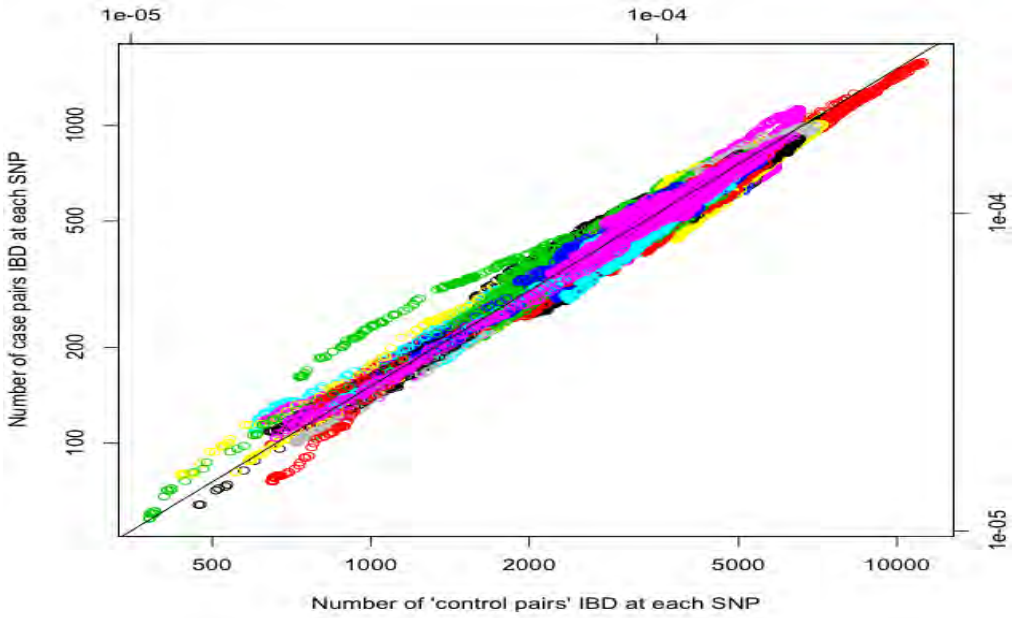
Results note 2.1. Table B. IBD pairs at each SNP in whole genome with the threshold of 3cM_1e-9

Chr	SNP	cM	bp	No. case_ case	No. case_ control	No. control_ control
1	rs4650610	105.2237858	79054556	627	2089	1937
1	rs1935562	81.0882568533	56976089	544	1933	1713
1	rs2176690	232.6536960334	219883931	598	2107	1987
…						
22	rs2285395	73.5750567738	49524956	104	355	306
Total: 274,735 SNPs						

* Case_case pairs were set as case pairs (*yi*); both case_control pairs (*ui*) and control_control pairs (*vi*) were set as "control pairs" (= *ui* + *vi)*.

Figure 2.2 is a scatterplot of case-pair sharing *yi* versus control-pair sharing *xi* as each of the 274,735 SNPs *i.* Using different colours to represent SNPs on different chromosomes, an outlier group of SNPs with relatively high case pair sharing on one chromosome stands out in green.

Figure 2.2. Plots of raw data of IBD with one point for each SNP. The green region clearly deviates from the black line, which indicates the proportion of case pairs in this region is higher than that of control pairs. The black line represents where the proportion of case pairs is equal to control pairs.



2.4.3. Fitting, testing Poisson model and converted to LOD scores

From examination of diagnostic plots (**Figure 2.3-2.5**), we found that the Poisson model provided the best fit to these data. **Figure 2.6** shows a plot of residuals from the Poisson model converted to LOD scores. The highest linkage signal, corresponding to the green outlier region in **Figure 2.2**, was observed on chromosome 19 with LOD=4.65 and $p=1.9\times10^{-6}$. As expected, a strong signal also was observed in the HLA region (LOD=3.58; **Figure 2.6**).

Figure 2.3. Fitting Poisson model for the IBD data. All the four real lines in these four modules fit well with the default lines, suggesting Poisson model is appropriate for this data. The residuals of the green region are higher than others.

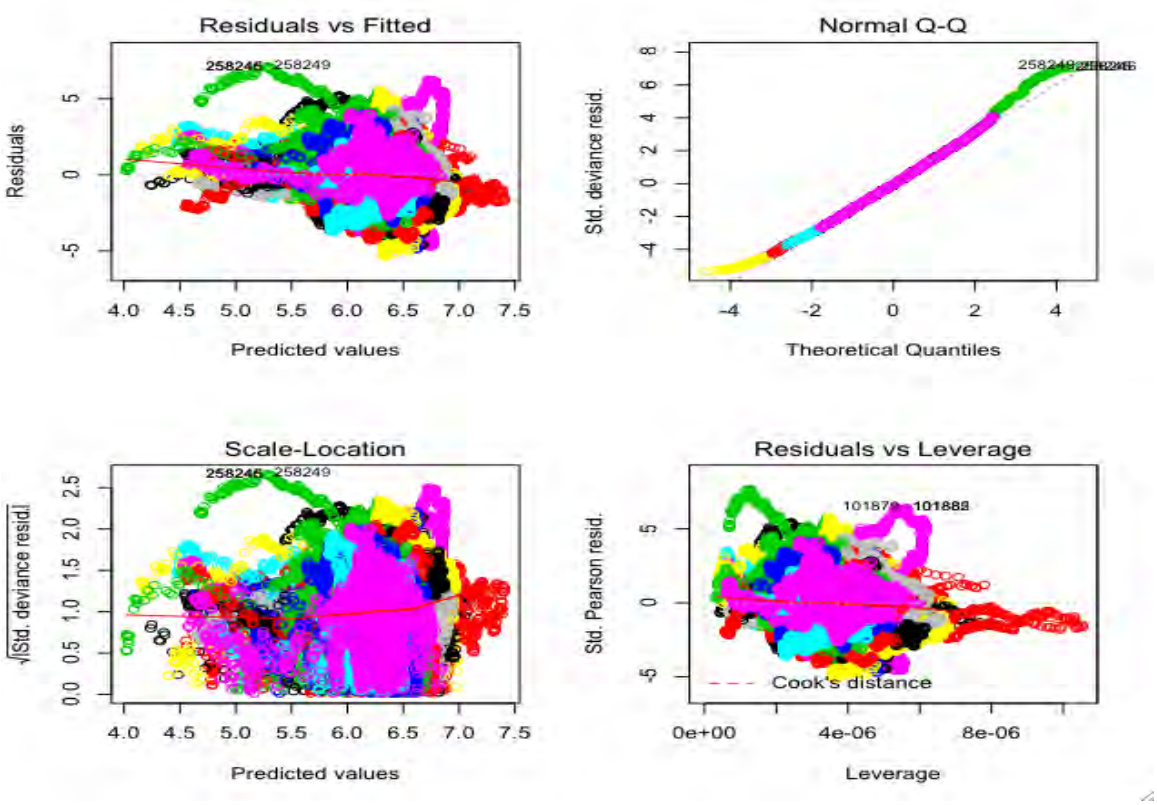


Figure 2.4. Fitting negative binomial model for the IBD data. All the four real lines in these four modules do not fit well with the default lines, suggesting negative binomial models are not suitable for this IBD data.

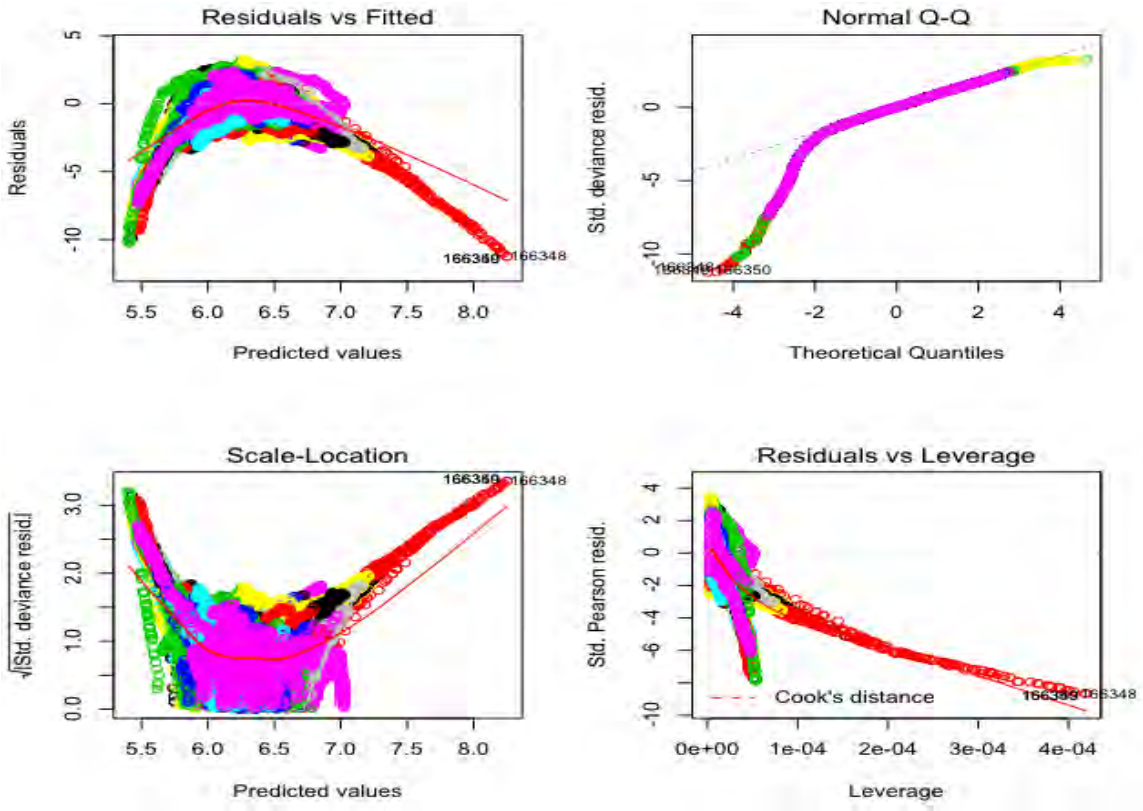


Figure 2.5. Fitting linear model for the IBD data. All the four real lines in these four modules do not fit well with the default lines, suggesting that linear models are not suitable for this IBD data.

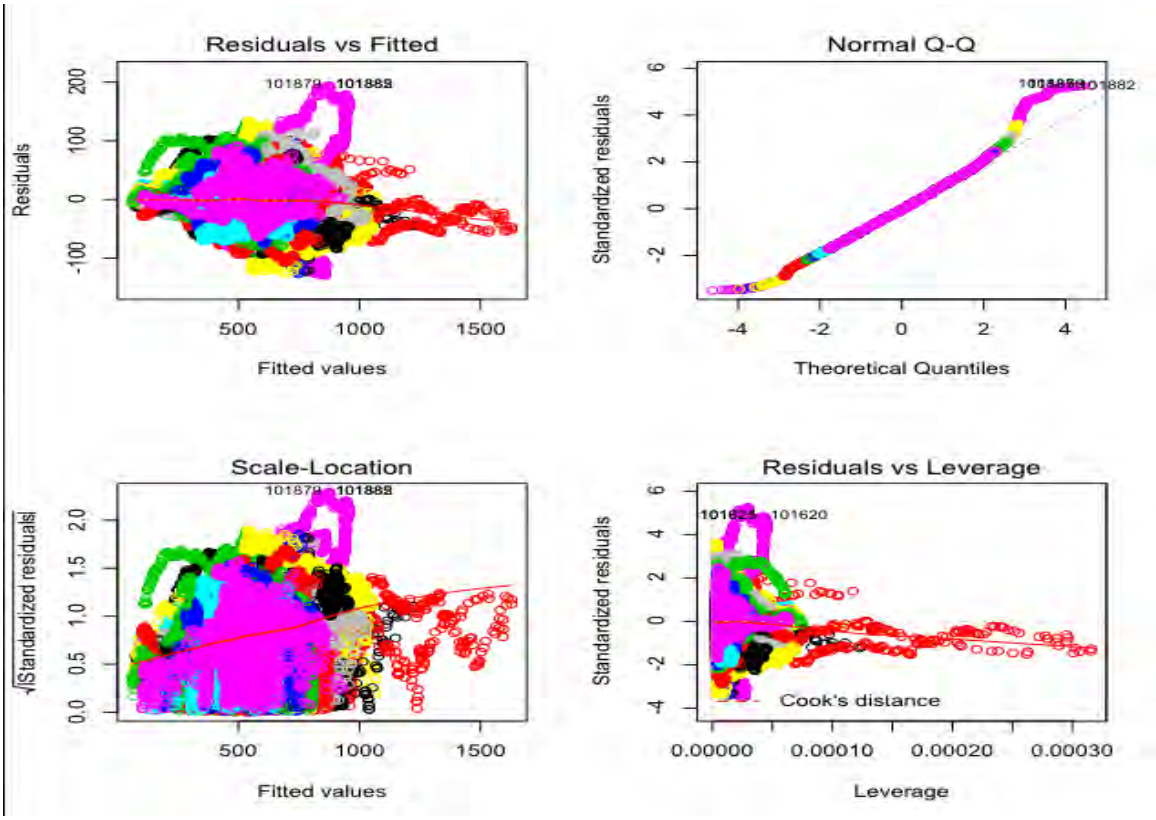
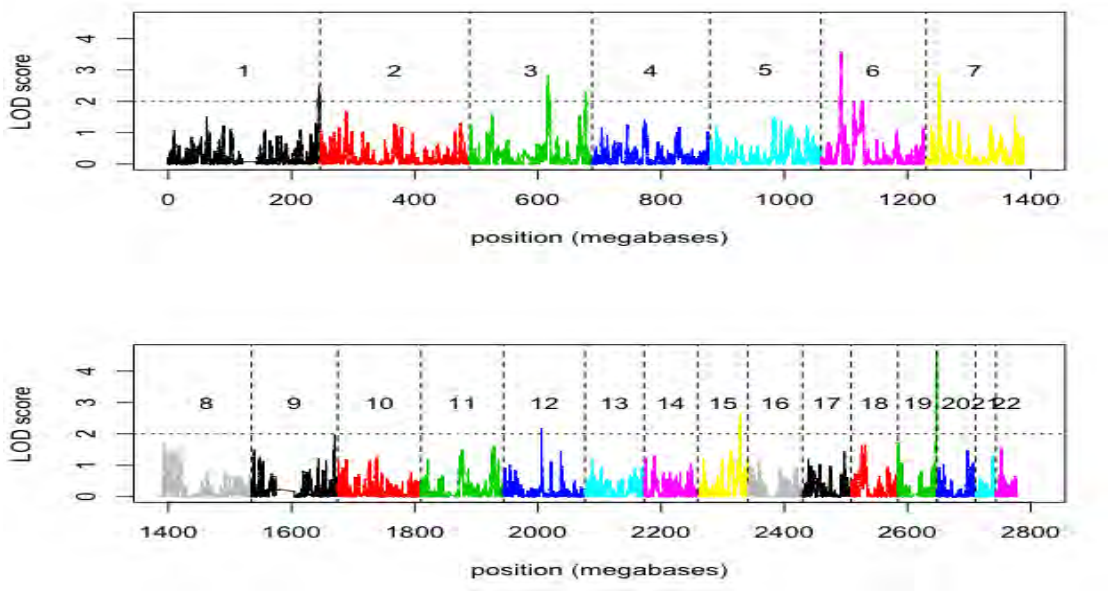


Figure 2.6. Plot of linkage scores along the whole genome with the IBD threshold of 3cM_1e-9 (shared haplotype segment >3cM and haplotype probability <10^{-9}). Chromosome 19 has the strongest linkage signal (LOD=4.65).



2.4.4 Analysis of linkage region on chromosome 19

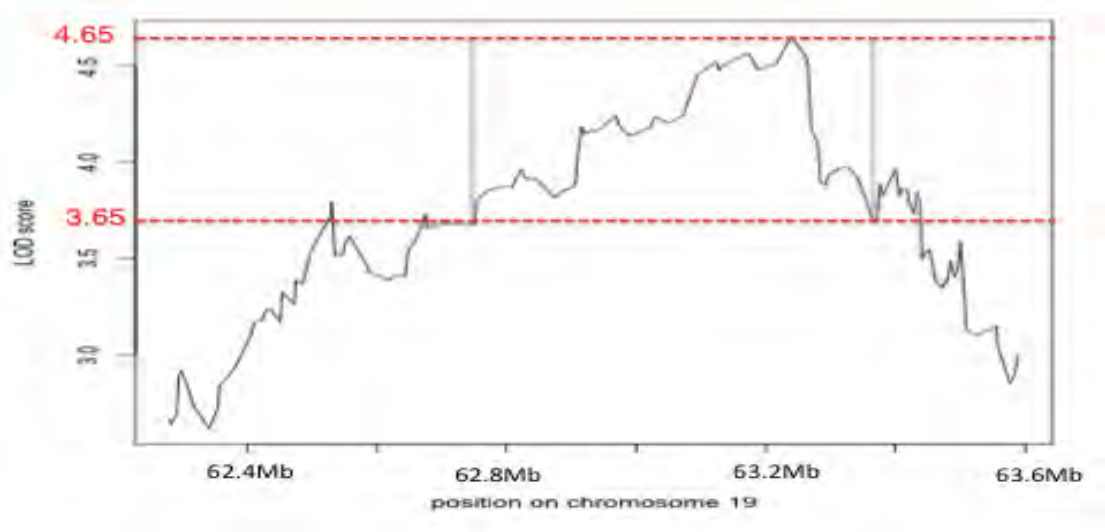
The linkage region on chromosome 19 with LOD score between 3.65 - LOD 4.65 is around 900kb in length (Hg18 chr19:62,529,738-63,437,743 bp) and corresponds to a cluster of zinc finger genes at 19q13.4, many of which have arisen by gene duplication. None of the genes in this region have been previously identified in published GWASs or associated with MS or autoimmune diseases.

Results note 2.2 The interested linkage region on chromosome 19

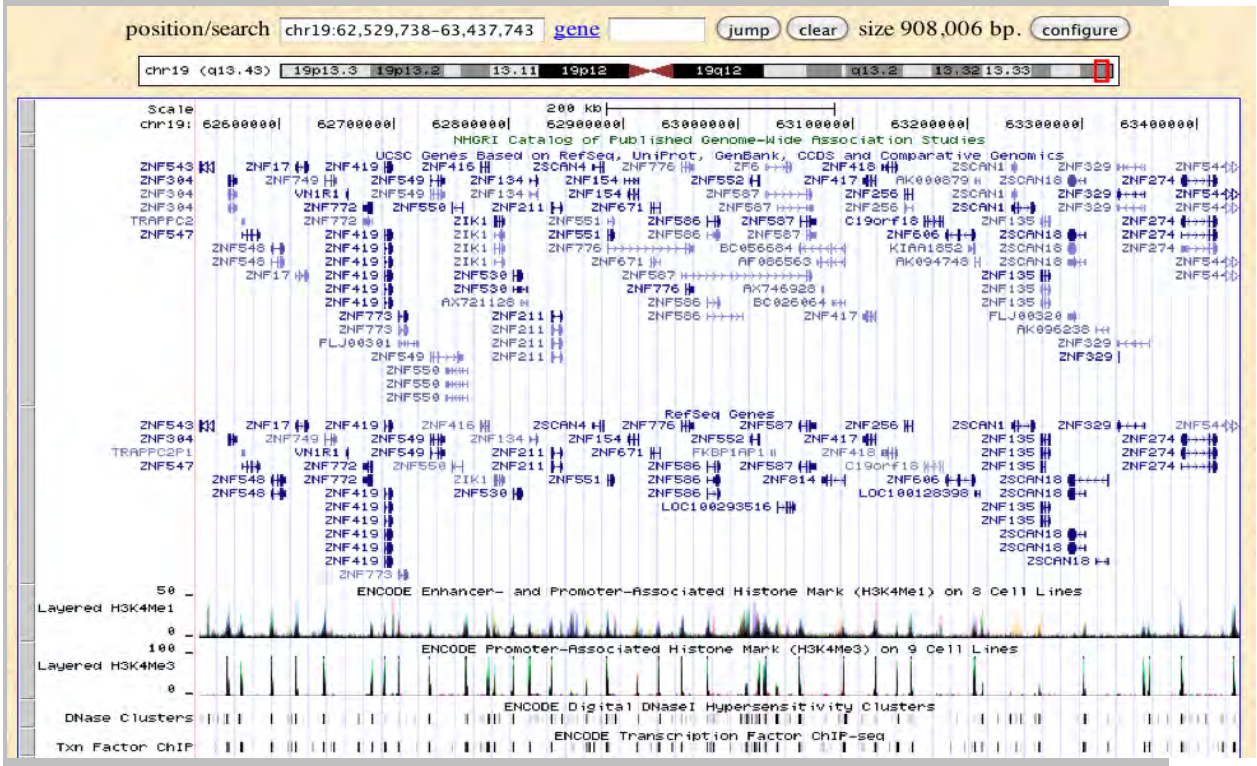
The interested linkage region generally refers to the region between the “highest LOD score” and “highest LOD score-1”. Results note 2.2. Figure A presents the interested

linkage region on chromosome 19, which is about 900 kb (between LOD 3.65 – LOD 4.65). Results note 2.2. Figure B presents the reference genes of the linkage region in UCSC genome browser. Most of the reference genes are zinc finger proteins and many of which have arisen by gene duplication.

Results note 2.2. Figure A. The interested linkage region on chromosome 19.



Results note 2.2. Figure B. RefSeq Genes in the interested linkage region on chromosome 19.



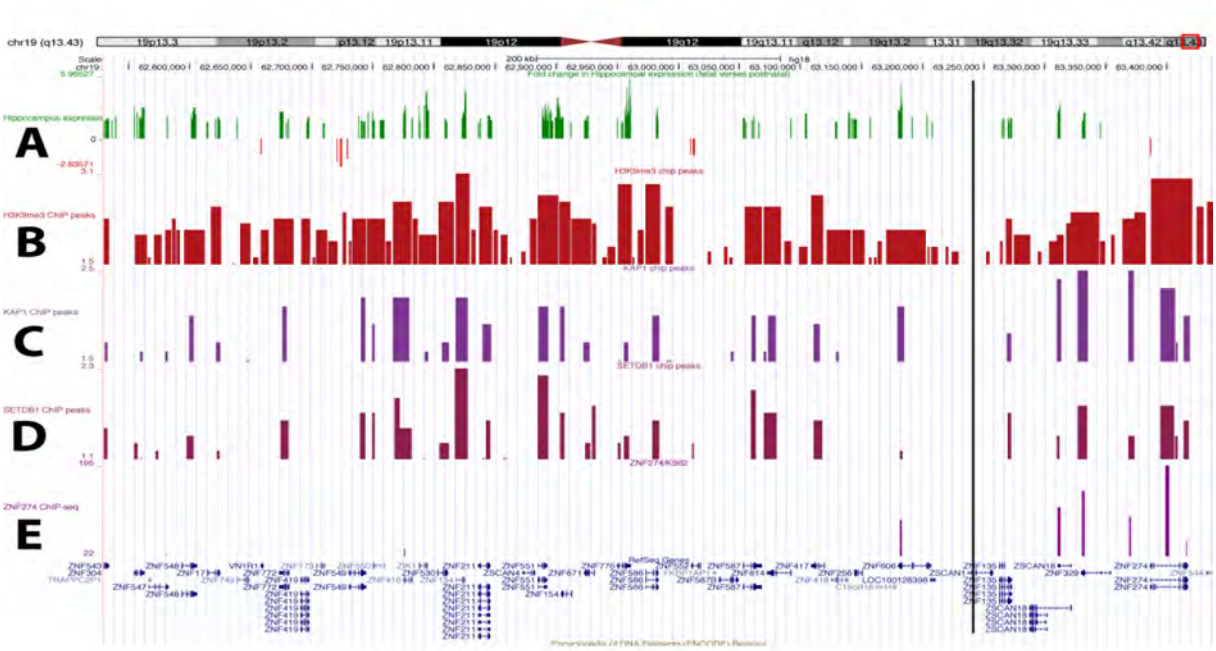
The genes in this region were examined to identify candidate genes with putative roles, which could, impact on susceptibility to MS. Published microarray expression data [37], profiling gene expression in the human hippocampus over a broad developmental range, were downloaded from Gene Expression Omnibus [38], series number GSE25219. Gene summary data was analysed in Partek Genomics Suite version 6.6 (Partek Inc., St. Louis, MO, USA) to generate expression profiles across all developmental periods for genes in the linkage region. Many of these genes exhibit similar expression profiles with high expression in early time points and low expression after birth. To categorize this observed trend, samples from a number of early foetal time periods (3, 4 and 5) (described in [37]) were grouped together and compared with expression of grouped samples from periods 9, 10, 11, and 12. Differential expression of exon level probe sets between these two groups was then analysed. The data points corresponding to individual probe sets, and

representing expression changes between these two developmental stages, were then aligned against the linkage region in a UCSC genome browser view. Differentially expressed probe sets were filtered using a false discovery rate adjusted p value cut off ($1.53x10^{-3}$) equivalent to a p value threshold of 0.01 and a fold change minimum 1.5. Those probe sets that passed this threshold were plotted on the UCSC browser screen view [39]. Genomic locations for Affymetrix exon level probe sets within the linkage region were downloaded from the UCSC table browser [40] and used to construct a bedGraph file of expression changes. The green bars indicate a higher expression in foetal time points compared to later time points. Some fold changes for genes in this region are very high (4-6 fold higher in foetal than post birth).

Although little is known about the majority of genes in this region, *ZNF274* is a DNA binding protein involved in regulation of H3K9me3 methylation at the 3' end of some ZNF genes by recruitment of the histone methyltransferase *SETDB1*, and the corepressor *TRIM28* (*KAP1*) [41]. To examine the pattern of H3K9me3 methylation in this region, genomic data on H3K9me3 methylation and *KAP1*, *SETDB1* and *ZNF274* binding in K562 cells [41] was used to make custom bedGraph files for visualization alongside the expression change with development (**Figure 2.7**). From this data we observed that the vast majority of genes in this region with high foetal expression levels are marked by both H3K9me3 methylation and bound by *KAP1* and *SETDB1* at the 3' end of the gene. A small number of genes are also bound at the 3' end of the transcript by *ZNF274*. We also observed a pattern in the level of H3K9me3 methylation, with two maximum levels at about position 62,850,000 and 63,400,000 and trailing off at position 63,250,00. This bimodal pattern also occurs in the *KAP1* and *SETDB1* binding data and is even more apparent when viewing a wider view of the region. This position, marked in the figure by

vertical black line, also corresponds with the position of rs159870 (chr19:63239261) and there is break in synteny with rodent genomes in this zone.

Figure 2.7. Screen shot from the UCSC genome browser illustrating expression regulation within the identified linkage region on Chromosome 19 (hg18) (*http://genome.ucsc.edu*). Human Refseq gene models are shown at the bottom of the figure. Custom bedGraph tracks illustrating expression regulation, as described in the manuscript, are shown. From top to bottom: (A) exon level expression fold change in hippocampus (FDR adjusted p value <0.01 and fold change >1.5) between early fetal (periods 3,4 and 5) and postnatal (periods 9,10,11 and 12) from Kang et al 2011, green bars indicate increased expression in fetal compared with postnatal and red bars indicate decreased expression in fetal compared with postnatal. ChIP-chip binding patterns of (B) H3K9me3 (C) TRIM28/KAP1 (D) SETDB1 and (E) ChIP-seq binding pattern of ZNF274 in K562 cells. For the ChIP-chip data log2 (ratio) values reflecting the ChIP enrichments are plotted on the Y axis. For the ChIP-seq data the number of tags reflecting the ChIP enrichments are plotted on the Y axis. ChIP-chip and ChIP-seq data are from Frietze et al 2010 supplementary data. Chromosomal coordinates and relative position on the chromosome is illustrated in the idogram at the top of the figure. The position of SNP rs159870 is shown by a vertical black line.



2.4.6 Comparison of IBD case pairs among different populations

We next examined patterns of IBD sharing at the SNP with the highest LOD score on chromosome 19 (rs159872). We compared the proportion of IBD case pairs in different populations to examine if there are some particular populations that contribute to more IBD case pairs in this locus, and found the Tasmanian population has the highest proportion of IBD case pairs, but this did not reach statistical significance when compared to other populations except for the US population (p-value=0.033). When compared with all other combined Australian populations, the Tasmanian population significantly contributed more IBD case pairs at this locus(p=0.004); and was significant when comparing with all other combined non-Tasmanian populations ($p=5.44\times10^{-5}$; **Table 2.2**).

Table 2.2: Comparison of IBD case pairs among different populations (rs159872 with the highest LOD score on chr19 (LOD=4.65))

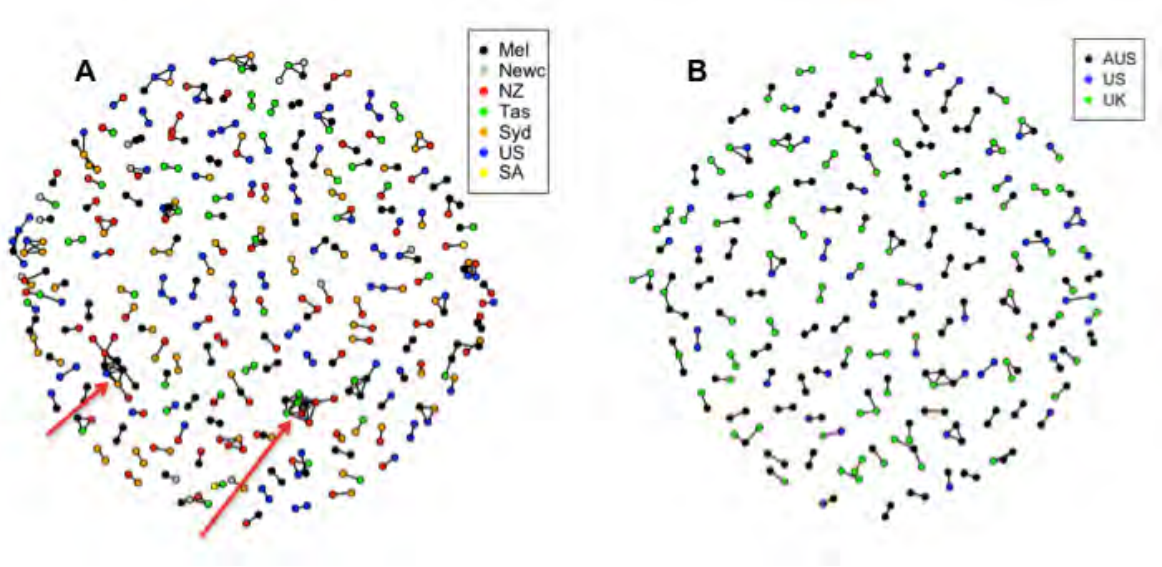
Population	No. case	No. IBD case pairs	% IBD case pairs	P-value
TAS	308	7	14.8×10^{-5}	Ref.
Mel	841	32	9.06×10^{-5}	0.216
Newc	111	0	0.00	1.00
Syd	541	14	9.58×10^{-5}	0.320
Other	32	0	0.00	1.00
AUS (non-TAS)	1525	46	3.96×10^{-5}	**0.004**
NZ	540	14	9.62×10^{-5}	0.321
US	879	22	5.70×10^{-5}	**0.033**
Non-TAS	2944	82	1.89×10^{-5}	$\mathbf{5.44\times10^{-5}}$

*% IBD case pairs=IBD pairs/case×(case-1)/2; (Fisher's Exact Test)

2.4.7 Networks of cases and controls sharing haplotypes on chromosome 19

Figure 2.8 shows networks of cases and controls sharing haplotypes IBD at the SNP with the highest LOD score on chromosome 19. The biggest cluster comprises 10 cases sharing a haplotype in which 4 cases come from Melbourne, 4 come from New Zealand, 1 come from Sydney and 1 come from USA. Another big cluster includes 2 cases from Melbourne, 3 cases from Tasmania and 4 cases from New Zealand (**Figure 2.8 A**) (red arrows). More generally, there are more big networks of cases sharing haplotypes than controls (**Figure 2.8 B**).

Figure 2.8. Networks of cases and controls sharing haplotypes IBD at the SNP (rs159872) with the highest LOD score on chromosome 19. A. Diagram showing networks of cases sharing haplotypes in common at the linkage region on chromosome 19. B. Diagram showing networks of controls sharing haplotypes in common at the linkage region on chromosome 19. Each dot represents an individual and each line connects pairs of individuals who share a haplotype. There are more big networks of cases sharing haplotypes than controls.



2.5 Discussion

We have applied BEAGLE fastIBD for the detection of rare MS variants utilising a large-scale GWAS dataset. We identified a high linkage signal on chromosome 19 with a p-value of 1.9×10^{-6} (LOD=4.65). In classical linkage analysis in small families, individuals are closely related and the segments of IBD tend to be fairly long (>10cM) which are easier to detect and less independent than IBD mapping, the generally-accepted threshold for genome-wide significance is $p=2.0\times10^{-5}$ [42]; while GWAS has more independent

tests than IBD mapping, the threshold of genome-wide significance is around $p=5.0\times10^{-8}$ [43], so the threshold of p-value for IBD mapping genome-wide significant should be between 5.0×10^{-8} and 2.0×10^{-5}. Recently, researchers demonstrated that the genome-wide significance thresholds for IBD mapping depend on the IBD segment size detected or IBD generations [32]. For example, an IBD segment size of 2 cM corresponds to 25 generations and the genome-wide significance threshold is 2.0×10^{-6}, while the segment size of 3.2 cM corresponds to 15 generations and the genome-wide significance threshold is 4.0×10^{-6} [32]. The strongest non-HLA linkage signal we detected in this study used a 3 cM segment size; which corresponds to 17 generations, thus the genome-wide significance threshold is between 4.0×10^{-6} and 2.0×10^{-6}. As such, the linkage signal on chromosome 19, with a p-value of 1.9×10^{-6}, was determined to be genome-wide significant.

2.5.1 Causal relationship between genes in linkage region and MS

Most genes in this linkage region are zinc finger (ZNF) proteins of which 32 genes have been suggested to be transcriptional regulators [44] (http://genome.ucsc.edu/). Seven genes (*ZNF134, ZNF135, ZNF154, ZNF549, ZNF606, ZNF671* and *ZSCAN1*) in this region belong to the Krüppel family of ZNF genes. Only a few ZNF genes in this region have known vertebrate homologues and it includes a number of primate specific KRAB-ZNF genes [45]. In humans KRAB-ZNF genes number about 400 and make up the largest group of C2H2 transcription factors [46] which are typically expressed at low levels and involved in cell specific silencing and driving different cell lineages.

Detailed analysis of genes in this region did not reveal any direct links with MS. However examination of their expression profiles in published data revealed a shared early

developmental CNS specific expression profile with 22 genes in this region being members of the expression module M20 described in [37], characterised by higher expression in all brain regions in early foetal time points followed by decreased expression prior to birth and very low expression thereafter. The M20 network of genes has a strong correlation with both neuronal differentiation and neuronal migration and a strong negative correlation with myelination [37].

Epigenetic mechanisms such as histone modification and DNA methylation are responsible for silencing many specific transcription factors including zinc finger genes, and the 3' end of many ZNF genes are specifically covered by H3K9me3 [47]. The zinc finger gene *ZNF274*, located within the linkage region, is involved in gene silencing through recruitment of the histone methytransferase complex TRIM28 (KAP1)/ SETDB1 to the 3' end of specific ZNF genes [41]. Examination of H3K9me3, KAP1 and SETDB1 binding data, confirms that many of the genes in the linkage region are covered by H3K9me3 at their 3'end (**Figure 2.7**). *ZNF274* also interacts with *p75NTR* and is predicted to play a role in programmed cell death during development [48]. A number of the genes in this area are also highly expressed in differentiated human neural cells compared to earlier stems cells (*ZNF549, ZNF324, ZNF548, ZNF264, ZNF671, ZSCAN1* and *ZSCAN18* are members of cluster A [49]). There is very little available evidence for involvement in immune cell activity for genes in this region. *ZNF304* is implicated in lymphocyte activation [50] and *ZNF274* has very high expression in activated eosinophils compared with other immune cell types [51]. Other genes in this region have relatively low expression and are not differentially regulated between immune cell types [51], as viewed in the immunological genome [52].

Together these findings suggest that many of the genes in this cluster may be involved in early differentiation of neuronal cells and potentially the silencing of genes required for myelination. Expression of ZNF genes is commonly detected in foetal brain and they are predicted to be involved in development of the nervous system, a KRAB zinc finger cluster on chromosome 8 has also been proposed to be involved in regulation of CNS development [53]. Although other clustered genes families have been shown to be co-expressed in cell types or tissues, previous studies have failed to identify coordinated expression of KRAB-ZNF gene clusters [45]. However earlier experiments did not examine the very early timepoints in CNS tissues included in the Kang dataset [37]. These expression profiles described in the M20 module are supported by two independent data set of both exon array level and RNA-seq expression data in early human CNS development available at the Allan Brain Atlas (http://developinghumanbrain.org/).

Thus this may be an example of a gene cluster of KRAB -ZNF genes exhibiting coordinated expression regulation, indicating the presence of a genomic regulatory block (GRB). Such regions are usually transcription factors controlled by highly conserved noncoding regions. Although the identification of GRBs remains difficult the evidence that we have collated is suggestive of two genomic regulatory blocks within the linkage region, interrupted at the position of SNP rs159870 where there is an absence of H3K9me3 methylation and a break in synteny (reviewed in [54]).

The underlying cause for susceptibility in this region could therefore be due, not to differences in a specific gene expression or protein product, but to differences in the tight expression regulation of a GRB. As mentioned above, many of the C2H2 zinc finger genes in this region have an expression profile consistent with silencing of genes required

for myelination. Further analysis needs to be undertaken to examine if these genes are co-regulated in demyelination and remyelination as well as CNS developmental states. Unfortunately, due to the species specificity of many of the KRAB-ZNF genes and the absence of rodent homologues of genes in this region, data from non-human models of demyelination and remyelination may not be useful.

Ideally, re-sequencing is the next step to refine this potential signal further. Unfortunately, resequencing of the region would be complicated since there are many gene duplications in this linkage region.

For the SNP (rs159872) with the highest LOD score on chromosome 19, we hypothesise that there are some difference between cases and controls sharing haplotypes in the linkage region among different populations. We found the Tasmanian MS population has the highest proportion of case IBD sharing, significantly higher than non-Tasmanian combined populations as well as other non-Tasmanian combined Australian populations. While Tasmania has the highest prevalence of MS in Australia, it is generally agreed that this is primarily driven by environmental effects related to, sunlight and/ or vitamin D [55]. However there is also a modest founder effect in Tasmania [56], which might result in an increase in MS susceptibility driven by rare variants IBD. Interestingly, we found there are more big networks of cases sharing haplotypes than controls, and one big case network comprises 3 Tasmanian cases, 4 New Zealand cases and 2 cases from Melbourne, which may indicate the potential causal variants or gene mutations exist in those big case networks. However, this SNP falls in a region of low/none methylation and correlates with a break in syteny, the significance of which is unclear.

2.5.2 Technical considerations

Even though Beagle fastIBD is several orders of magnitude faster than Beagle IBD, IBD analysis remains moderately computationally intensive on a dataset of this size (8,977 individuals and 274,735 SNPs). For instance, on chromosome 2 with 22,607 SNPs, the computation time for each run was approximately 4.6 hours with memory requirement of 3.3 GB on 2 cores of a SGI Altix ICE 8200 HPC cluster computer node.

However, we also found IBD analysis limitations: it is only suited to discover rare variants if all variants act in the same direction in one gene. For example, the identified rare variants in *BRCA1* and *BRCA2* gene all increase risk of breast cancer [57], and the four rare variants identified in *IFIH1* gene all protect against type I diabetes [19]. If some rare variants increase risk while others in the same gene decrease risk then the signal in the region will be attenuated. In addition, we found IBD analysis is very sensitive to genotyping error, resulting in reducing signal strength. The linkage signal detected depends on a lot of markers or long haplotypes, containing up to hundreds of SNPs, a single error occurring in reading a single marker significantly reduces the signal. In our data, samples came from different GWAS using different genotyping chips in different locations, which at least in part, may decrease the potential signal strength from our study. Furthermore, resequencing would be complicated by gene duplication and repeat regions, since the linkage region detected in this study had many gene duplicates, thus replication in other independent dataset is needed.

The optimal method to detect rare disease-causing variants is whole genome sequencing of thousands of samples. When this becomes affordable, there will remain a role for IBD

analysis to prioritize regions for follow-up analysis and minimize the massive multiple testing burden. Just as linkage analysis is now used to identify regions for follow-up in whole genome sequencing and exome sequencing of Mendelian disease families, and linkage analysis can be used to weight regions for GWA analysis [27].

In summary, we have applied IBD analysis to a large complex disease GWA dataset and identified a linkage signal with genome-wide significance, although it. While our most significant result is of equivocal significance, and lies in a region that is hard to validate via sequencing, we believe IBD analysis has considerable potential, particularly to help interpret whole-genome sequencing data in complex trait studies.

2.6 Summary

Genome-wide association studies have identified about 110 common variants associated with MS, but these loci only explain a fraction of the heritability of MS. Some missing heritability may be caused by rare variants that have been suggested to play an important role in the aetiology of complex diseases such as MS. However current genetic and statistical methods for detecting rare variants are expensive and time consuming. 'Population-based linkage analysis' (PBLA) or so called identity-by-descent (IBD) mapping is a novel way to detect rare variants in extant GWAS datasets.

We employed BEAGLE fastIBD to search for rare MS variants utilising IBD mapping in a large GWAS dataset of 3,543 cases and 5,898 controls. We identified a genome-wide significant linkage signal on chromosome 19 (LOD=4.65; $p=1.9\times10^{-6}$). Network analysis of cases and controls sharing haplotypes on chromosome 19 further strengthened the association as there are more large networks of cases sharing haplotypes than controls.

This linkage region includes a cluster of zinc finger genes of unknown function. Analysis of genome wide transcriptome data suggests that genes in this zinc finger cluster may be involved in very early developmental regulation of the CNS. Our study also indicates that BEAGLE fastIBD allowed identification of rare variants in large unrelated population with moderate computationally intensity. Even with the development of whole-genome sequencing, IBD mapping still may be a promising way to narrow down the region of interest for sequencing priority.

2.7 Postscript

The chapter has described how to use IBD mapping to detect rare variants, and we have identified that a genome-wide significant linkage signal was on chromosome 19, which includes a cluster of zinc finger protein genes that may be involved in very early developmental regulation of the CNS. In the next chapter I will discuss how to detect rare variants/and or disease-causing variants of MS by using another powerful method - family-based whole genome sequencing analysis.

2.8 References

1. Sadovnick AD, Armstrong H, Rice GP, Bulman D, Hashimoto L, et al. (1993) A population-based study of multiple sclerosis in twins: update. Ann Neurol 33: 281-285.
2. Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, et al. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 11: 504-510.
3. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
4. Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, et al. (2009) Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet 17: 1309-1313.
5. Ban M, McCauley JL, Zuvich R, Baker A, Bergamaschi L, et al. (2010) A non-synonymous SNP within membrane metalloendopeptidase-like 1 (MMEL1) is associated with multiple sclerosis. Genes Immun 11: 660-664.

6. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767-778.
7. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41: 776-782.
8. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851-862.
9. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2009) CD226 Gly307Ser association with multiple autoimmune diseases. Genes Immun 10: 5-10.
10. International Multiple Sclerosis Genetics Consortium (IMSGC). (2010) Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 19: 953-962.
11. Jakkula E, Leppa V, Sulonen AM, Varilo T, Kallio S, et al. (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am J Hum Genet 86: 285-291.
12. Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, et al. (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 42: 495-497.
13. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
14. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet.
15. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, et al. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection. Am J Hum Genet 90: 720-726.
16. Jersild C, Dupont B, Fog T, Platz PJ, Svejgaard A (1975) Histocompatibility determinants in multiple sclerosis. Transplant Rev 22: 148-163.
17. Jersild C, Fog T (1972) Histocompatibility (HL-A) antigens associated with multiple sclerosis. Acta neurologica Scandinavica Supplementum 51: 377.
18. O'Gorman C, Lin R, Stankovich J, Broadley SA (2012) Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data. Neuroepidemiology 40: 1-12.
19. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387-389.
20. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519-525.
21. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592-599.
22. Yu Z, Zhu Y, Chen-Plotkin AS, Clay-Falcone D, McCluskey L, et al. (2011) PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLoS One 6: e17951.

23. Kang SJ, Scott WK, Li YJ, Hauser MA, van der Walt JM, et al. (2006) Family-based case-control study of MAOA and MAOB polymorphisms in Parkinson disease. Mov Disord 21: 2175-2180.
24. Browning SR (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178: 2123-2132.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
26. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.
27. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using Linkage Genome Scans to Improve Power of Association in Genome Scans. Am J Hum Genet 78: 243-252.
28. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19: 318-326.
29. Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86: 526-539.
30. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. Am J Hum Genet 88: 173-182.
31. Browning SR (2006) Multilocus Association Mapping Using Variable-Length Markov Chains. Am J Hum Genet 78: 903-913.
32. Browning SR, Thompson EA (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190: 1521-1531.
33. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, et al. (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. Am J Hum Genet 85: 750-755.
34. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
35. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.
36. R Development Core Team. (2010) R: A language and envirnment for statistical computing. Vienana, Australia: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org
37. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. Nature 478: 483-489.
38. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 39: D1005-1010.
39. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
40. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493-496.
41. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. PLoS One 5: e15082.
42. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11: 241-247.

43. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32: 227-234.
44. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. Genome Biol 4: 231.
45. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res 16: 669-677.
46. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252-263.
47. O'Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, et al. (2007) Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. PLoS Genet 3: e89.
48. Casademunt E, Carter BD, Benzel I, Frade JM, Dechant G, et al. (1999) The zinc finger protein NRIF interacts with the neurotrophin receptor p75(NTR) and participates in programmed cell death. Embo J 18: 6050-6061.
49. Fathi A, Hatami M, Hajihosseini V, Fattahi F, Kiani S, et al. (2011) Comprehensive gene expression analysis of human embryonic stem cells during differentiation into neural cells. PLoS One 6: e22856.
50. Sabater L, Ashhab Y, Caro P, Kolkowski EC, Pujol-Borrell R, et al. (2002) Identification of a KRAB-containing zinc finger protein, ZNF304, by AU-motif-directed display method and initial characterization in lymphocyte activation. Biochem Biophys Res Commun 293: 1066-1072.
51. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, et al. (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat Immunol 7: 274-283.
52. Heng TS, Painter MW (2008) The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol 9: 1091-1094.
53. Lorenz P, Dietmann S, Wilhelm T, Koczan D, Autran S, et al. (2010) The ancient mammalian KRAB zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of C2H2 zinc finger evolution associated with unique expression profiles in human tissues. BMC Genomics 11: 206.
54. Navratilova P, Becker TS (2009) Genomic regulatory blocks in vertebrates and implications in human disease. Brief Funct Genomic Proteomic 8: 333-342.
55. Simpson S, Jr., Blizzard L, Otahal P, Van der Mei I, Taylor B (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. J Neurol Neurosurg Psychiatry 82: 1132-1141.
56. Rubio JP, Bahlo M, Butzkueven H, van Der Mei IA, Sale MM, et al. (2002) Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. Am J Hum Genet 70: 1125-1137.
57. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72: 1117-1130.

Supplementary R commands

1) Fitting and testing model for IBD data

```
library(MASS)
m1 <- glm.nb(case ~ control)
#nb = negative binomial regression
par(mfrow=c(2,2))
plot(m1,col=d$chr)

m2 <- lm(case ~ control)
#lm = linear regression
par(mfrow=c(2,2))
plot(m2,col=d$chr)

m3 <- glm(case ~ offset(log(control)),family=poisson)
#glm = poisson regression
par(mfrow=c(2,2))
plot(m3,col=d$chr)
```

2) Plot of residuals from the Poisson model converted to LOD scores

```
par(mfrow=c(2,1))
c <- 1
shift <- -(min(d$bp[d$chr==c]))/1000000
plot(d$bp[d$chr==c]/1000000+shift,LOD[d$chr==c],type="l",ylim=c(0,max(LOD)),xlim=c(0,14
 00),xlab="position (megabases)",ylab="LOD score",col=d$chr[d$chr==c])
lines(c(-100,1500),c(2,2),lty=3)
mid <- (min(d$bp[d$chr==c])+max(d$bp[d$chr==c]))/2000000+shift
text(mid,3,c)
for (c in 2:7) {
   shift <- shift+(max(d$bp[d$chr==c-1])-min(d$bp[d$chr==c]))/1000000
   lines(rep(shift+min(d$bp[d$chr==c])/1000000,2),c(-10,60),lty=2)
   lines(d$bp[d$chr==c]/1000000+shift,LOD[d$chr==c],col=d$chr[d$chr==c])
   mid <- (min(d$bp[d$chr==c])+max(d$b[d$chr==c]))/2000000+shift
   text(mid,3,c)
}
c <- 8
shift <- shift+(max(d$bp[d$chr==c-1])-min(d$bp[d$chr==c]))/1000000
plot(d$bp[d$chr==c]/1000000+shift,LOD[d$chr==c],type="l",ylim=c(0,max(LOD)),xlim=c(1400
 ,2800),xlab="position (megabases)",ylab="LOD score",col=d$chr[d$chr==c])
lines(c(1300,3000),c(2,2),lty=3)
mid <- (min(d$bp[d$chr==c])+max(d$bp[d$chr==c]))/2000000+shift
text(mid,3,c)
for (c in 9:22) {
   shift <- shift+(max(d$bp[d$chr==c-1])-min(d$bp[d$chr==c]))/1000000
   lines(rep(shift+min(d$bp[d$chr==c])/1000000,2),c(-10,60),lty=2)
   lines(d$bp[d$chr==c]/1000000+shift,LOD[d$chr==c],col=d$chr[d$chr==c])
   mid <- (min(d$bp[d$chr==c])+max(d$bp[d$chr==c]))/2000000+shift
   text(mid,3,c)
}
```

3) Network analysis

```
library(network)
n <- read.table("/Users/rlin2/Documents/BEAGLE/population_groups/3cM_1e-
 9_chr19.network",header=FALSE)
groups <- read.table("/Users/rlin2/Documents/BEAGLE/population_groups/3cM_1e-
 9_chr19.grouplist",header=FALSE)
m <- as.matrix(n)
g <- network(m,directed=FALSE)
a <- plot.network(g,vertex.col=groups[,2],vertex.cex=0.7)
xmax <- max(a[,1])
ymax <- max(a[,2])
legend(xmax-
10,ymax+10,legend=c("Mel","Newc","NZ","Tas","Syd","US","SA"),pch=rep(16,7),col=1:7)
```

Appendix 2A: Publication in Chapter 2

Rui Lin, Jac Charlesworth, Jim Stankovich, Victoria M. Perreau, Matthew A. Brown, ANZgene Consortium, Bruce V Taylor. Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis. *PLoS One*. 2013;8(3):e56379.

OPEN ACCESS Freely available online

PLOS ONE

Identity-by-Descent Mapping to Detect Rare Variants Conferring Susceptibility to Multiple Sclerosis

Rui Lin[1], Jac Charlesworth[1], Jim Stankovich[1,2], Victoria M. Perreau[3], Matthew A. Brown[4], ANZgene Consortium[¶], Bruce V. Taylor[1]*

1 Menzies Research Institute Tasmania, University of Tasmania, Hobart, Australia, **2** Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, **3** Centre for Neurosciences, Department of Anatomy and Neuroscience, University of Melbourne, Melbourne, Australia, **4** University of Queensland Diamantina Institute, Brisbane, Australia

Abstract

Genome-wide association studies (GWAS) have identified around 60 common variants associated with multiple sclerosis (MS), but these loci only explain a fraction of the heritability of MS. Some missing heritability may be caused by rare variants that have been suggested to play an important role in the aetiology of complex diseases such as MS. However current genetic and statistical methods for detecting rare variants are expensive and time consuming. 'Population-based linkage analysis' (PBLA) or so called identity-by-descent (IBD) mapping is a novel way to detect rare variants in extant GWAS datasets. We employed BEAGLE fastIBD to search for rare MS variants utilising IBD mapping in a large GWAS dataset of 3,543 cases and 5,898 controls. We identified a genome-wide significant linkage signal on chromosome 19 (LOD = 4.65; $p = 1.9 \times 10^{-6}$). Network analysis of cases and controls sharing haplotypes on chromosome 19 further strengthened the association as there are more large networks of cases sharing haplotypes than controls. This linkage region includes a cluster of zinc finger genes of unknown function. Analysis of genome wide transcriptome data suggests that genes in this zinc finger cluster may be involved in very early developmental regulation of the CNS. Our study also indicates that BEAGLE fastIBD allowed identification of rare variants in large unrelated population with moderate computational intensity. Even with the development of whole-genome sequencing, IBD mapping still may be a promising way to narrow down the region of interest for sequencing priority.

Citation: Lin R, Charlesworth J, Stankovich J, Perreau VM, Brown MA, et al. (2013) Identity-by-Descent Mapping to Detect Rare Variants Conferring Susceptibility to Multiple Sclerosis. PLoS ONE 8(3): e56379. doi:10.1371/journal.pone.0056379

Editor: Amanda Ewart Toland, Ohio State University Medical Center, United States of America

Received November 7, 2012; **Accepted** January 8, 2013; **Published** March 5, 2013

Funding: This work was supported by the National Health and Medical Research Council (NHMRC, Australia) [Project 605511]; and the Australian Research Council [LP110100473]. MAB was supported by an NHMRC Senior Principal Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Bruce.Taylor@utas.edu.au

¶ Membership of the Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) is provided in the Acknowledgments.

Introduction

Multiple sclerosis (MS) is a complex neurological disease of the central nervous system (CNS) triggered by environmental and genetic factors. There is considerable evidence for a significant genetic component to MS susceptibility, such as a higher concordance rate in monozygotic twins (24%–30%) than dizygotic twins (3%–5%) [1,2]. As for other immune diseases, genome-wide association studies (GWAS) have been highly successful for MS: uncovering around 60 common genetic variants associated with disease [3–13]. The majority of these variants lie near genes with known functions in the immune system and these variants have also been associated with other autoimmune diseases, often in the opposite direction [14]. Virtually all of the variants confer modest increases in disease risk, the outstanding exception being the strong association with the *HLA-DRB1*15:01* allele in the major histocompatibility complex (MHC), which was first detected in the 1970's [15,16].

Despite this success, the variants identified by GWAS to date only explain 18–24% of the heritability of MS [13,17]. While much of the missing heritability is probably explained by common variants of even smaller effect sizes, some heritability may be explained by rare variants of larger effect size. Standard analysis of GWAS data is not designed to detect associations with rare variants that many believe may play an important role in the aetiology of complex traits [18–20]. Interestingly, GWAS have had less success for putative neurodegenerative diseases, such as Parkinson's disease, than for MS. For these diseases, family-based approaches detecting rare variants have been more successful [21,22]. This raises the possibility that rare variants under negative selection pressure are relatively more important in the genetic architecture of neurodegenerative processes, whereas common variants under balancing selection are more important in the genetic architecture of immunological processes. Discovery of rare MS susceptibility variants may alter perspectives on the relative importance of immunological & neurodegenerative processes in MS onset.

Standard analyses of GWAS data are not designed to detect associations with low frequency variants (MAF≤5%), and other strategies are required. One approach is to re-sequence loci containing common susceptibility variants identified from GWAS studies. This strategy was used to detect rare variants in *IFIH1* conferring protection to type I diabetes [18]. However this strategy

Table 1. Sample numbers from GWAS (after cleaning).

GWAS dataset	Country of origin		No. Case	No. Control	Total	No. SNPs
	Cases	Controls				
ANZgene [3]	AUS, NZ	US, UK	1,608	3,404	5,012	300,900
WTCCC2 [13]	AUS	-	766	-	766	586,393
QIMR [32]	-	AUS	-	1,516	1,516	529,292
GeneMSA [6]	US	US	878	805	1,683	550,677
Total			3,252	5,725	8,977	274,735*

*The number of SNPs that passed QC in all 4 GWAS datasets.
AUS = Australia, NZ = New Zealand.
doi:10.1371/journal.pone.0056379.t001

precludes the identification of new loci. Eventually it will be possible to overcome this limitation by whole genome sequencing, but it remains prohibitively expensive to perform adequately-powered studies. An alternative is to re-analyse GWAS data using identity-by-descent (IBD) mapping [23], also referred to as 'population-based linkage analysis' (PBLA) [24]. PBLA describes linkage analysis applied at the population level to detect mega base-scale regions where cases have inherited long haplotypes from distant ancestors, 10–100 generations ago. IBD mapping is performed on the unrelated individuals to determine whether these mega-base scale regions are identical and inherited from a common ancestor. If the common ancestor lived more than ten generations ago the individuals will share very short tracts of genetic material, and a shared haplotype that is very rare is also very likely to be IBD. HapMap Phase 3 identified that lower frequency variants should, on average, be younger than more common variants; and thus display a greater extent of haplotype sharing [25]. Therefore, if case pairs can be detected with long shared haplotypes (generally one to five megabases) inherited from distant common ancestors, then rare variants influencing disease risk can be localised. Even when whole genome sequencing becomes cheap enough to pursue with substantial sample sizes, IBD mapping may still help reduce the massive multiple testing problem by prioritizing regions. This is similar to the technique of prioritising association signals in regions of linkage [26].

Several methods of IBD mapping have been published: these include PLINK [24], GERMLINE [27], BEAGLE IBD [28] and BEAGLE fastIBD [29]. The models employed by PLINK and GERMLINE assume SNPs are in linkage equilibrium, and so 'pruning' of SNPs [24] is required to avoid false positives due to under-estimates of population haplotype frequencies. However pruning of SNPs in incomplete linkage disequilibrium (LD) discards potentially useful information and reduces power. BEAGLE IBD and fastIBD implement a variable length Hidden Markov Model [30] to account for LD and model haplotype frequencies more accurately. BEAGLE fastIBD runs considerably faster than BEAGLE IBD (of the order of 1000 times faster with large GWAS datasets). This is mainly because 1) it does not formally model IBD status ('IBD'/'not IBD') between pairs of individuals using a Hidden Markov Model as in BEAGLE IBD; 2) it stores haplotype frequencies in a data dictionary (as in GERMLINE) which means computational time scales with sample size n like n log n instead of n^2.

To detect MS rare variants, we here use BEAGLE fastIBD to perform an IBD analysis on several large MS GWAS datasets comprised 3543 cases and 5898 controls. We identified a region of high significance on chromosome 19q13.43, with a genome-wide

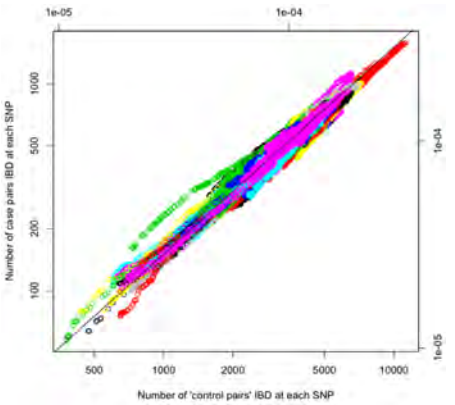


Figure 1. Plots of raw data of IBD with one point for each SNP. The green region is obvious outstanding from the black line, which indicates the proportion of case pairs in this region higher than that of control pairs. The black line represents where the proportion of case pairs equal to control pairs.
doi:10.1371/journal.pone.0056379.g001

significant localisation signal ($p = 1.9 \times 10^{-6}$; LOD = 4.65) using thresholds based on IBD segment length greater than 3 cM and the probability p-value less than 10^{-9} (3cM_1e-9). This locus was deemed genome-wide significant according to the recently established genome-wide significance thresholds set for IBD mapping [31]. Analysis of expression data and investigation of genes in this area support the hypothesis for regulation of gene expression in this region to impact upon development or health of CNS tissue. Our analyses also illustrate some of the practical issues to deal with in IBD analyses, and demonstrate that IBD mapping can form a potentially powerful method for detecting rare variants in unrelated individuals at the population level.

Methods

Study subjects

All the MS cases and controls were recruited and genotyped from MS GWAS totaling 3,543 cases and 5,898 controls. Of these, 1,618 cases and 3,413 controls were from an Australian and New Zealand MS GWAS conducted by the Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) [3], and those DNA samples were genotyped on the Illumina Infinium Hap370CNV array [3]. An additional 861 Australian and New Zealand MS cases were genotyped with the Illumina Human660-Quad chip as part of a GWAS performed by the International MS Genetics Consortium (IMSGC) and the Wellcome Trust Case Control Consortium-2 (WTCCC2) [13]. Controls included 1,531 unrelated Australian samples from a GWAS genotyped by Queensland Institute of Medical Research (QIMR) with the Illumina Human610-Quad chip [32], and 1064 MS cases and 954 controls genotyped with the Sentrix® HumanHap550 BeadChip from a GWAS conducted in the US (GeneMSA) [6] [accessed via dbGAP].

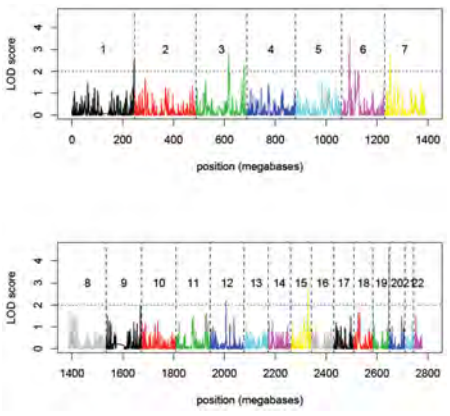


Figure 2. Plot of linkage scores along the whole genome with the IBD threshold of 3cM_1e-9 (shared haplotype segment >3 cM and haplotype probability $p<10^{-9}$). Chromosome 19 has the strongest linkage signal (LOD = 4.65).
doi:10.1371/journal.pone.0056379.g002

Quality control of data

Conservative quality control measures were imposed both on the individual datasets before merging, and in the combined dataset after merging: SNPs with call rates less than 0.95 or in Hardy-Weinberg disequilibrium ($p<10^{-7}$) were discarded, as were samples with call rates less than 0.98. Duplicates and close relatives were also removed. This data cleaning was performed using PLINK.

A principal components analysis (PCA) was conducted by EIGENSTRAT [33] to exclude ancestry outliers and examine population structure within the remaining samples. First, SNPs in strong LD were pruned (using the PLINK – indep command with options 50 5 1.5), and then we excluded previously identified regions of high LD [34]. Outliers in the PCA were excluded using standard settings in Eigenstrat (more than six standard deviations from the mean along the first 10 principal components). All chromosomal locations refer to Human genome version hg18.

Running BEAGLE fastIBD and results processing

The fastIBD analyses were conducted using BEAGLE (http://faculty.washington.edu/browning/beagle/). In brief, genotypes for the merged, cleaned dataset were converted to BEAGLE format by using the linkage2Beagle.jar utility program. We then used the BEAGLE method for phasing the data and identifying IBD segments simultaneously, using the 'fastibdthreshold' option. This procedure was run 10 times for each chromosome starting with different seeds of the random number generator.

The output of these calculations was a series of "putative" IBD segments shared between pairs of individuals. Each segment comes with the following information attached: ids for the pair of individuals, first and last SNPs in the IBD segment, length of the segment in centimorgans, and probability of the two individuals both carrying the segment if it was not IBD. We filtered these segments using various maximum probabilities and minimum segment lengths, as recommended in the BEAGLE manual. Results from the 10 runs were combined by taking the union of IBD segments detected in each run. From the final list of segments, we wrote a Perl script to count numbers of case-case pairs (yi), case-control pairs (ui) and control-control pairs (vi) estimated to share haplotypes IBD at each SNP.

Analysis of IBD

We focused on the detection of loci where groups of cases have inherited rare susceptibility alleles IBD. To do this, we modelled IBD sharing yi in case-case pairs ("case pairs") as a function of IBD sharing in $xi = ui + vi$ in case-control pairs and control-control pairs combined ("control pairs").

We tried various methods to model the yi as a function of the xi: linear regression, negative binomial regression and Poisson regression. Models were fitted using R [35] and goodness of fit was assessed by examining diagnostic plots (**SR_commands S1**).

At SNPs i with more IBD sharing in cases than expected, residuals zi from the fitted models should be large and positive. To present residuals on a scale more familiar to geneticists, we converted them to LOD scores using the formula $LODi = zi^2/(2*\log_e(10))$.

At the SNPs with the highest LOD scores, we calculated the proportions of case pairs sharing IBD in various populations, and plotted networks of case and control pairs sharing IBD with each other using the R network package (http://cran.rproject.org/web/packages/network/index.html).

Results

Study samples from GWAS after cleaning

11 individuals were excluded due to call rates less than 0.98 and an additional 202 individuals were excluded because they were close relatives or duplicates. PCA was conducted on a subset of 77,856 SNPs not in LD, which were common to all sample sets. Through successive iterations 251 outliers (37 AUS cases; 9 AUS controls; 6 UK controls; 97 US cases; 102 US controls) were excluded. All datasets overlapped well after the removal of outliers (**Figure S1**). In summary, following cleaning there were 3,243 cases and 5,725 controls with 274,735 autosomal SNPs in the final analysis (**Table 1**).

Results of IBD analysis

We detected IBD with the threshold of IBD segment greater than 3 cM and the haplotype probability p-value less than 10^{-9} (3cM_1e-9). A strong linkage signal was observed in the HLA region (LOD = 3.58), while the strongest signal in non-HLA region was on chromosome 19 (LOD = 4.65), which reached genome-wide significance according to the recent established genome-wide significance threshold set for IBD mapping [31].

Figure 1 is a scatterplot of case-pair sharing yi versus control-pair sharing xi as each of the 274,735 SNPs i. Using different colours to represent SNPs on different chromosomes, an outlier group of SNPs with relatively high case pair sharing on one chromosome stands out in green.

Fitting, testing Poisson model and converting to LOD scores

From examination of diagnostic plots (**Figure S2, S3, S4**), we found that the Poisson model provided the best fit for these data. **Figure 2** shows a plot of residuals from the Poisson model converted to LOD scores. The highest linkage signal, corresponding to the green outlier region in **Figure 1**, was observed on chromosome 19 with LOD = 4.65 and $p = 1.9\times10^{-6}$. As expected, a strong signal also was observed in the HLA region (LOD = 3.58; **Fig. 1**).

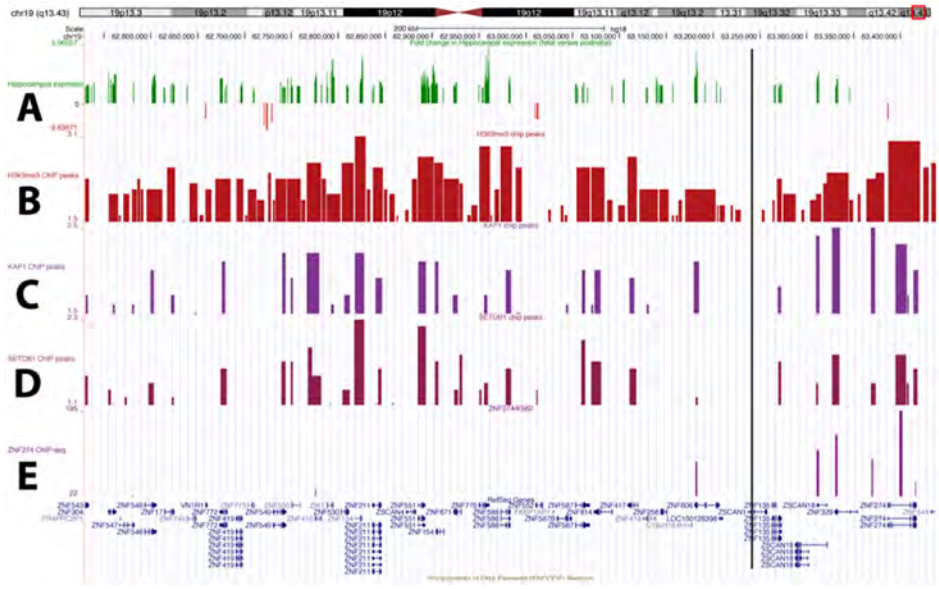


Figure 3. Screen shot from the UCSC genome browser illustrating expression regulation within the identified linkage region on Chromosome 19 (hg18) (*http://genome.ucsc.edu*). Human Refseq gene models are shown at the bottom of the figure. Custom bedGraph tracks illustrating expression regulation, as described in the manuscript, are shown. From top to bottom: (A) exon level expression fold change in hippocampus (FDR adjusted p value <0.01 and fold change >1.5) between early fetal (periods 3,4 and 5) and postnatal (periods 9,10,11 and 12) from Kang et al 2011, green bars indicate increased expression in fetal compared with postnatal and red bars indicate decreased expression in fetal compared with postnatal. ChIP-chip binding patterns of (B) H3K9me3 (C) TRIM28/KAP1 (D) SETDB1 and (E) ChIP-seq binding pattern of ZNF274 in K562 cells. For the ChIP-chip data log2 (ratio) values reflecting the ChIP enrichments are plotted on the Y axis. For the ChIP-seq data the number of tags reflecting the ChIP enrichments are plotted on the Y axis. ChIP-chip and ChIP-seq data are from Frietze et al 2010 supplementary data. Chromosomal coordinates and relative position on the chromosome is illustrated in the idogram at the top of the figure. The position of SNP rs159870 is shown by a vertical black line.
doi:10.1371/journal.pone.0056379.g003

Analysis of Linkage region on chromosome 19

The linkage region on chromosome 19 with LOD scores between 3.65 and 4.65 is around 900kb in length (Hg18 chr19: 62,529,738–63,437,743 bp) and corresponds to a cluster of zinc finger genes at 19q13.4, many of which have arisen by gene duplication. None of the genes in this region have been previously identified in published GWAS or associated with MS or autoimmune diseases.

The genes in this region were examined to identify candidate genes with putative roles, which could, impact on susceptibility to MS. Published microarray expression data [36], profiling gene expression in the human hippocampus over a broad developmental range, were downloaded from Gene Expression Omnibus [37], series number GSE25219. Gene summary data was analysed in Partek Genomics Suite version 6.6 (Partek Inc., St. Louis, MO, USA) to generate expression profiles across all developmental periods for genes in the linkage region. Many of these genes exhibit similar expression profiles with high expression in early time points and low expression after birth. To categorize this observed trend, samples from a number of early foetal time periods (3, 4 and 5) (described in [36]) were grouped together and compared with expression of grouped samples from periods 9, 10, 11, and 12. Differential expression of exon level probe sets between these two groups was then analysed. The data points corresponding to individual probe sets, and representing expression changes between these two developmental stages, were then aligned against the linkage region in a UCSC genome browser view. Differentially expressed probe sets were filtered using a false discovery rate adjusted p value cut off (1.53×10^{-3}) equivalent to a p value threshold of 0.01 and a fold change minimum 1.5. Those probe sets that passed this threshold were plotted on the UCSC browser screen view [38]. Genomic locations for Affymetrix exon level probe sets within the linkage region were downloaded from the UCSC table browser [39] and used to construct a bedGraph file of expression changes. The green bars indicate a higher expression in foetal time points compared to later time points. Some fold changes for genes in this region are very high (4–6 fold higher in foetal than post birth).

Although little is known about the majority of genes in this region, *ZNF274* is a DNA binding protein involved in regulation of H3K9me3 methylation at the 3′ end of some ZNF genes by recruitment of the histone methyltransferase *SETDB1*, and the corepressor *TRIM28* (*KAP1*) [40]. To examine the pattern of H3K9me3 methylation in this region, genomic data on H3K9me3 methylation and *KAP1*, *SETDB1* and *ZNF274* binding in K562 cells [40] was used to make custom bedGraph files for visualization

alongside the expression change with development (**Fig. 3**). From this data we observed that the vast majority of genes in this region with high foetal expression levels are marked by both H3K9me3 methylation and bound by *KAP1* and *SETDB1* at the 3′ end of the gene. A small number of genes are also bound at the 3′ end of the transcript by *ZNF274*. We also observed a pattern in the level of H3K9me3 methylation, with two maximum levels at about position 62,850,000 and 63,400,000 and trailing off at position 63,250,00. This bimodal pattern also occurs in the *KAP1* and *SETDB1* binding data and is even more apparent when viewing a wider view of the region. This position, marked in the figure by vertical black line, also corresponds with the position of rs159870 (chr19: 63239261) and there is break in synteny with rodent genomes in this zone.

Comparison of IBD sharing among different populations

We next examined patterns of IBD sharing at the SNP with the highest LOD score on chromosome 19 (rs159872). We compared the proportion of IBD case pairs in different populations to determine whether there are particular populations that contribute to more IBD case pairs at this locus; and found the Tasmanian population has the highest proportion of IBD case pairs. When compared with all other combined Australian populations, the Tasmanian population significantly contributed more IBD case pairs at this locus ($p = 0.004$); and was significant when compared with all other combined non-Tasmanian populations ($p = 5.44\times10^{-5}$; **Table 2**).

Networks of cases and controls sharing haplotypes on chromosome 19

Figure 4 shows networks of cases and controls sharing haplotypes IBD at the SNP with the highest LOD score on chromosome 19. The biggest cluster comprises 10 cases sharing a haplotype in which 4 cases were from Melbourne, 4 from New Zealand, 1 from Sydney and 1 from USA. Another big cluster includes 2 cases from Melbourne, 3 cases from Tasmania and 4 cases from New Zealand (**Fig. 4A**). More generally, there are more networks of cases sharing haplotypes than controls.

Table 2. Comparison of IBD case pairs among different populations (rs159872 with the highest LOD score on chr19; LOD = 4.65).

Population	No. case	No. IBD case pairs	% IBD case pairs	p-value
TAS	308	7	14.80×10^{-5}	Ref.
Mel	841	32	9.06×10^{-5}	0.22
Newc	111	0	0.00	1.00
Syd	541	14	9.58×10^{-5}	0.32
Other	32	0	0.00	1.00
AUS (non-TAS)	1525	46	3.96×10^{-5}	0.004
NZ	540	14	9.62×10^{-5}	0.32
US	879	22	5.70×10^{-5}	0.033
Non-TAS	2944	82	1.89×10^{-5}	5.44×10^{-5}

*% IBD case pairs = IBD pairs/case×(case-1)/2; (Fisher's Exact Test).

doi:10.1371/journal.pone.0056379.t002

Discussion

We have applied BEAGLE fastIBD for the detection of rare MS variants utilising a large-scale GWAS dataset. We identified a high linkage signal on chromosome 19 with a p-value of 1.9×10^{-6} (LOD = 4.65). In classical linkage analysis in small families, individuals are closely related and the segments of IBD tend to be fairly long (>10 cM) which are easier to detect and less independent than IBD mapping, the generally-accepted threshold for genome-wide significance is $p = 2.0\times10^{-5}$ [41]; while GWAS has more independent tests than IBD mapping, the threshold of genome-wide significance is around $p = 5.0\times10^{-8}$ [42], so the threshold of p-value for IBD mapping genome-wide significant should be between 5.0×10^{-8} and 2.0×10^{-5}. Recently, researchers demonstrated that the genome-wide significance thresholds for IBD mapping depend on the IBD segment size detected or IBD generations [31]. For example, an IBD segment size of 2 cM corresponds to 25 generations and the genome-wide significance threshold is 2.0×10^{-6}, while the segment size of 3.2 cM corresponds to 15 generations and the genome-wide significance threshold is 4.0×10^{-6} [31]. The strongest non-HLA linkage signal we detected in this study used a 3 cM segment size; which corresponds to 17 generations, thus the genome-wide significance threshold is between 4.0×10^{-6} and 2.0×10^{-6}. As such, the linkage signal on chromosome 19, with a p-value of 1.9×10^{-6}, was determined to be genome-wide significant.

Causal relationship between genes in linkage region and MS

Most genes in this linkage region are zinc finger (ZNF) proteins of which 32 genes have been suggested to be transcriptional regulators [43] (http://genome.ucsc.edu/). Seven genes (*ZNF134, ZNF135, ZNF154, ZNF549, ZNF606, ZNF671* and *ZSCAN1*) in this region belong to the Krüppel family of ZNF genes. Only a few ZNF genes in this region have known vertebrate homologues and it includes a number of primate specific KRAB-ZNF genes [44]. In humans KRAB-ZNF genes number about 400 and make up the largest group of C2H2 transcription factors [45] which are typically expressed at low levels and involved in cell specific silencing and driving different cell lineages.

Detailed analysis of genes in this region did not reveal any direct links with MS. However examination of their expression profiles in published data revealed a shared early developmental CNS specific expression profile with 22 genes in this region being members of the expression module M20 described in [36], characterised by higher expression in all brain regions in early foetal time points followed by decreased expression prior to birth and very low expression thereafter. The M20 network of genes has a strong correlation with both neuronal differentiation and neuronal migration and a strong negative correlation with myelination [36].

Epigenetic mechanisms such as histone modification and DNA methylation are responsible for silencing many specific transcription factors including zinc finger genes, and the 3′ end of many ZNF genes are specifically covered by H3K9me3 [46]. The zinc finger gene *ZNF274*, located within the linkage region, is involved in gene silencing through recruitment of the histone methytransferase complex TRIM28 (KAP1)/SETDB1 to the 3′ end of specific ZNF genes [40]. Examination of H3K9me3, KAP1 and SETDB1 binding data, confirms that many of the genes in the linkage region are covered by H3K9me3 at their 3′end (**Fig. 3**). *ZNF274* also interacts with *p75NTR* and is predicted to play a role in programmed cell death during development [47]. A number of the genes in this area are also highly expressed in differentiated

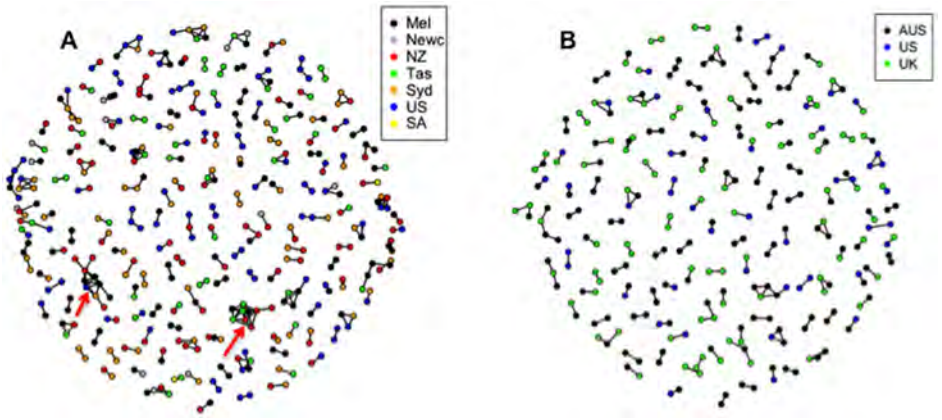


Figure 4. Networks of cases and controls sharing haplotypes IBD at the SNP with the highest LOD score on chromosome 19. (**A**) Networks of cases sharing haplotypes in common at the linkage region on chromosome 19. (**B**) Networks of controls sharing haplotypes in common at the linkage region on chromosome 19. Each dot represents an individual and each line connects pairs of individuals who share a haplotype. There are more big networks of cases sharing haplotypes than controls.
doi:10.1371/journal.pone.0056379.g004

human neural cells compared to earlier stems cells (*ZNF549, ZNF324, ZNF548, ZNF264, ZNF671, ZSCAN1* and *ZSCAN18* are members of cluster A [48]). There is very little available evidence for involvement in immune cell activity for genes in this region. *ZNF304* is implicated in lymphocyte activation [49] and *ZNF274* has very high expression in activated eosinophils compared with other immune cell types [50]. Other genes in this region have relatively low expression and are not differentially regulated between immune cell types [50], as viewed in the immunological genome [51].

Together these findings suggest that many of the genes in this cluster may be involved in early differentiation of neuronal cells and potentially the silencing of genes required for myelination. Expression of ZNF genes is commonly detected in foetal brain and they are predicted to be involved in development of the nervous system, a KRAB zinc finger cluster on chromosome 8 has also been proposed to be involved in regulation of CNS development [52]. Although other clustered genes families have been shown to be co-expressed in cell types or tissues, previous studies have failed to identify coordinated expression of KRAB-ZNF gene clusters [44]. However earlier experiments did not examined the very early timepoints in CNS tissues included in the Kang dataset [36]. These expression profiles described in the M20 module are supported by two independent data set of both exon array level and RNA-seq expression data in early human CNS development available at the Allan Brain Atlas (http://developinghumanbrain.org/).

Thus this may be an example of a gene cluster of KRAB -ZNF genes exhibiting coordinated expression regulation, indicating the presence of a genomic regulatory block (GRB). Such regions are usually transcription factors controlled by highly conserved noncoding regions. Although the identification of GRBs remains difficult the evidence that we have collated is suggestive of two genomic regulatory blocks within the linkage region, interrupted at the position of SNP rs159870 where there is an absence of H3K9me3 methylation and a break in synteny (reviewed in [53]).

The underlying cause for susceptibility in this region could therefore be due, not to differences in a specific gene expression or protein product, but to differences in the tight expression regulation of a GRB. As mentioned above, many of the C2H2 zinc finger genes in this region have an expression profile consistent with silencing of genes required for myelination. Further analysis needs to be undertaken to examine if these genes are co-regulated in demyelination and remyelination as well as CNS developmental states. Unfortunately, due to the species specificity of many of the KRAB-ZNF genes and the absence of rodent homologues of genes in this region, data from non-human models of demyelination and remyelination may not be useful.

Ideally, re-sequencing is the next step to refine this potential signal further. Unfortunately, resequencing of the region would be complicated since there are many gene duplications in this linkage region.

For the SNP (rs159872) with the highest LOD score on chromosome 19, we hypothesise that there are some difference between cases and controls sharing haplotypes in the linkage region among different populations. We found the Tasmanian MS population has the highest proportion of case IBD sharing, significantly higher than non-Tasmanian combined populations as well as other non-Tasmanian combined Australian populations. While Tasmania has the highest prevalence of MS in Australia, it is generally agreed that this is primarily driven by environmental effects related to, sunlight and/or vitamin D [54]. However there is also a modest founder effect in Tasmania [55], which might result in an increase in MS susceptibility driven by rare variants IBD. Interestingly, we found there are more big networks of cases sharing haplotypes than controls, and one big case network comprises 3 Tasmanian cases, 4 New Zealand cases and 2 cases from Melbourne, which may indicate the potential causal variants or gene mutations exist in those big case networks. However, this

SNP falls in a region of low/none methylation and correlates with a break in syteny, the significance of which is unclear.

Technical considerations

Even though Beagle fastIBD is several orders of magnitude faster than Beagle IBD, IBD analysis remains moderately computationally intensive on a dataset of this size (8,977 individuals and 274,735 SNPs). For instance, on chromosome 2 with 22,607 SNPs, the computation time for each run was approximately 4.6 hours with memory requirement of 3.3 GB on 2 cores of a SGI Altix ICE 8200 HPC cluster computer node.

However, we also found IBD analysis limitations: it is only suited to discover rare variants if all variants act in the same direction in one gene. For example, the identified rare variants in *BRCA1* and *BRCA2* gene all increase risk of breast cancer [56], and the four rare variants identified in *IFIH1* gene all protect against type I diabetes [18]. If some rare variants increase risk while others in the same gene decrease risk then the signal in the region will be attenuated. In addition, we found IBD analysis is very sensitive to genotyping error, resulting in reducing signal strength. The linkage signal detected depends on a lot of markers or long haplotypes, containing up to hundreds of SNPs, a single error occurring in reading a single marker significantly reduces the signal. In our data, samples came from different GWAS using different genotyping chips in different locations, which at least in part, may decrease the potential signal strength from our study. Furthermore, resequencing would be complicated by gene duplication and repeat regions, since the linkage region detected in this study had many gene duplicates, thus replication in other independent dataset is needed.

The optimal method to detect rare disease-causing variants is whole genome sequencing of thousands of samples. When this becomes affordable, there will remain a role for IBD analysis to prioritize regions for follow-up analysis and minimize the massive multiple testing burden. Just as linkage analysis is now used to identify regions for follow-up in whole genome sequencing and exome sequencing of Mendelian disease families, and linkage analysis can be used to weight regions for GWA analysis [26].

In summary, we have applied IBD analysis to a large complex disease GWA dataset and identified a linkage signal with genome-wide significance, although it. While our most significant result is of equivocal significance, and lies in a region that is hard to validate via sequencing, we believe IBD analysis has considerable potential, particularly to help interpret whole-genome sequencing data in complex trait studies.

Supporting Information

Figure S1 Principal components analysis for the dataset. Most individuals in the dataset are of predominantly northern European ancestry (right hand side), but some have southern European ancestry (left hand side) (one dot for each individual).
(TIF)

Figure S2 Fitting Poisson model for the IBD data. All the four real lines in these four modules fit well with the default lines, suggesting Poisson model is appropriate for this data. The residuals of the green region are higher than others.
(TIF)

Figure S3 Fitting negative binomial model for the IBD data. All the four real lines in these four modules fit not well with the default lines, suggesting negative binomial model is not suitable for this IBD data.
(TIF)

Figure S4 Fitting linear model for the IBD data. All the four real lines in these four modules fit not well with the default lines, suggesting linear model is not suitable for this IBD data.
(TIF)

SR_commands S1 1) Fitting and testing model for IBD data. 2) Plot of residuals from the Poisson model converted to LOD scores. 3) Network analysis.
(PDF)

Acknowledgments

We thank IMSGC and WTCCC2, QIMR and the ANZgene Consortium for access to the data relating to the MS associated SNPs. We thank the Tasmanian Partnership for Advanced Computing (TPAC) for access to the high performance computing facilities at the University of Tasmania.

The members of the ANZgene Consortium are: Alan Baxter (School of Pharmacy and Molecular Sciences, James Cook University, Townsville, Australia), Allan Kermode (Department of Neurology, Sir Charles Gairdner Hospital, Nedlands, Australia), Melanie Bahlo (Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia), William Carroll (Department of Neurology, Sir Charles Gairdner Hospital, Nedlands, Australia), Helmut Butzkueven (Department of Medicine, Royal Melbourne Hospital, Parkville, Australia), David Booth (Westmead Millennium Institute, University of Sydney, Sydney, Australia), Graeme Stewart (Westmead Millennium Institute, University of Sydney, Sydney, Australia), James Wiley (Howard Florey Institute, University of Melbourne, Melbourne, Australia), Judith Field (Howard Florey Institute, University of Melbourne, Melbourne, Australia), Lotti Tajouri (Genomics Research Centre, Griffith University, Brisbane, Australia), Lyn Griffiths (Genomics Research Centre, Griffith University, Brisbane, Australia), Michael Barnett (Brain and Mind Research Institute, University of Sydney, Camperdown, Australia), Pablo Moscato (Hunter Medical Research Institute, Newcastle, Australia), Robert Heard (Westmead Millennium Institute, University of Sydney, Sydney, Australia), Rodney Scott (School of Biomedical Sciences, University of Newcastle, Newcastle, Australia), Shaun McColl (School of Molecular & Biomedical Science, University of Adelaide, Adelaide, Australia), Simon Foote (Australian School of Advanced Medicine, Macquarie University, Sydney, Australia), Simon Broadley (School of Medicine, Griffith University, Gold Coast Campus, Australia), Mark Slee (School of Medicine, Flinders University of South Australia, Adelaide, Australia), Steve Vucic (Western Clinical School, University of Sydney, Randwick, Australia), Trevor Kilpatrick (Centre for Neurosciences, Department of Anatomy and Neuroscience, University of Melbourne, Melbourne, Australia).

Author Contributions

Conceived and designed the experiments: JS. Performed the experiments: MAB ANZgene Consortium. Analyzed the data: RL JS VMP. Contributed reagents/materials/analysis tools: ANZgene Consortium. Wrote the paper: RL JS VMP JC BVT.

References

1. Sadovnick AD, Armstrong H, Rice GP, Bulman D, Hashimoto L, et al. (1993) A population-based study of multiple sclerosis in twins: update. Ann Neurol 33: 281–285.
2. Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, et al. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 11: 504–510.
3. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824–828.
4. Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, et al. (2009) Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet 17: 1309–1313.

5. Ban M, McCauley JL, Zuvich R, Baker A, Bergamaschi L, et al. (2010) A non-synonymous SNP within membrane metalloendopeptidase-like 1 (MMEL1) is associated with multiple sclerosis. Genes Immun 11: 660–664.
6. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767–778.
7. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41: 776–782.
8. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851–862.
9. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2009) CD226 Gly307Ser association with multiple autoimmune diseases. Genes Immun 10: 5–10.
10. The International Multiple Sclerosis Genetics Consortium (IMSGC). (2010) Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 19: 953–962.
11. Jakkula E, Leppa V, Sulonen AM, Varilo T, Kallio S, et al. (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am J Hum Genet 86: 285–291.
12. Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, et al. (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 42: 495–497.
13. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214–219.
14. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, et al. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection. Am J Hum Genet 90: 720–726.
15. Jersild C, Dupont B, Fog T, Platz PJ, Svejgaard A (1975) Histocompatibility determinants in multiple sclerosis. Transplant Rev 22: 148–163.
16. Jersild C, Fog T (1972) Histocompatibility (HL-A) antigens associated with multiple sclerosis. Acta neurologica Scandinavica Supplementum 51: 377.
17. O'Gorman C, Lin R, Stankovich J, Broadley SA (2012) Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data. Neuroepidemiology 40: 1–12.
18. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.
19. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519–525.
20. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592–599.
21. Yu Z, Zhu Y, Chen-Plotkin AS, Clay-Falcone D, McCluskey L, et al. (2011) PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLoS One 6: e17951.
22. Kang SJ, Scott WK, Li YJ, Hauser MA, van der Walt JM, et al. (2006) Family-based case-control study of MAOA and MAOB polymorphisms in Parkinson disease. Mov Disord 21: 2175–2180.
23. Browning SR (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178: 2123–2132.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
25. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
26. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using Linkage Genome Scans to Improve Power of Association in Genome Scans. Am J Hum Genet 78: 243–252.
27. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19: 318–326.
28. Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86: 526–539.
29. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. Am J Hum Genet 88: 173–182.
30. Browning SR (2006) Multilocus Association Mapping Using Variable-Length Markov Chains. Am J Hum Genet 78: 903–913.
31. Browning SR, Thompson EA (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190: 1521–1531.
32. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, et al. (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. Am J Hum Genet 85: 750–755.
33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
34. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944–947.
35. R Development Core Team (2010) R: A language and envirmment for statistical computing. Vienana, Australia: R Foundation for Statistical Computing. Available: http://www.R-project.org.
36. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. Nature 478: 483–489.
37. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets – 10 years on. Nucleic Acids Res 39: D1005–1010.
38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
39. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493–496.
40. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. PLoS One 5: e15082.
41. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11: 241–247.
42. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32: 227–234.
43. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. Genome Biol 4: 231.
44. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res 16: 669–677.
45. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252–263.
46. O'Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, et al. (2007) Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. PLoS Genet 3: e89.
47. Casademunt E, Carter BD, Benzel I, Frade JM, Dechant G, et al. (1999) The zinc finger protein NRIF interacts with the neurotrophin receptor p75(NTR) and participates in programmed cell death. Embo J 18: 6050–6061.
48. Fathi A, Hatami M, Hajihosseini V, Fattahi F, Kiani S, et al. (2011) Comprehensive gene expression analysis of human embryonic stem cells during differentiation into neural cells. PLoS One 6: e22856.
49. Sabater L, Ashhab Y, Caro P, Kolkowski EC, Pujol-Borrell R, et al. (2002) Identification of a KRAB-containing zinc finger protein, ZNF304, by AU-motif-directed display method and initial characterization in lymphocyte activation. Biochem Biophys Res Commun 293: 1066–1072.
50. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, et al. (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. Nat Immunol 7: 274–283.
51. Heng TS, Painter MW (2008) The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol 9: 1091–1094.
52. Lorenz P, Dietmann S, Wilhelm T, Koczan D, Autran S, et al. (2010) The ancient mammalian KRAB zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of C2H2 zinc finger evolution associated with unique expression profiles in human tissues. BMC Genomics 11: 206.
53. Navratilova P, Becker TS (2009) Genomic regulatory blocks in vertebrates and implications in human disease. Brief Funct Genomic Proteomic 8: 333–342.
54. Simpson S Jr., Blizzard L, Otahal P, Van der Mei I, Taylor B (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. J Neurol Neurosurg Psychiatry 82: 1132–1141.
55. Rubio JP, Bahlo M, Butzkueven H, van Der Mei IA, Sale MM, et al. (2002) Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. Am J Hum Genet 70: 1125–1137.
56. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 72: 1117–1130.

Chapter 3. Detecting disease-causing variants of MS by using family-based whole genome sequencing analysis

3.1 Preface

IBD mapping, as described in Chapter 2, is one method of localising regions likely to harbor disease relevant rare variation. However the best method of identifying rare or private genetic variants is whole-genome sequencing (WGS) of disease enriched families. At present, WGS remains prohibitively expensive to perform adequately-powered case-control studies, but family-based WGS is a powerful alternative. In this chapter I will describe the first stages of disease-causing variant detection using pedigree based WGS. This chapter is the first component of a much larger project, the bulk of which remains outside the scope of this dissertation. The analysis of pedigree based WGS for the identification of rare, disease-relevant variants can form the basis of a major research project in itself; however for the purposes of my dissertation I will focus on the initial variant filtering and confirmation of most likely risk variants within this chapter. Future analyses of these data are outside the limits of this thesis but I will retain a significant interest and input into the further analysis of this family.

3.2 Introduction

Multiple sclerosis is a complex disease driven by environmental and genetic factors. Studies have demonstrated a significant genetic component to MS susceptibility, such as a

higher concordance rate in monozygotic twins (24%-30%) than dizygotic twins (3%-5%) [1,2], and more than 100 common genetic variants outside the HLA region have been identified as associated with MS by GWASs [3-14]. However, rare, disease-causing variants of MS have been little identified [15]. Next-generation sequencing of whole-genomes has proven to be a valuable method for the discovery of the genetic underpinnings of rare and complex diseases [16]. Although cheaper than Sanger sequencing (over the entire genome), whole-genome sequencing remains prohibitively expensive to perform adequately-powered population-based studies. Pedigree based studies allow the identification of "private" disease relevant variants and by their very structure and the number of transmission events in a large pedigree they are also enriched for rare-variant identification. In addition, identification of inheritance patterns in the pedigree permits the detection of ~70% of sequencing errors (resulting in >99.999% accuracy), and sharply reduces the search space for disease-causing variants [17]. However, previous family-based researches showed little success in the search for rare/disease-causing variants in MS [18-21], due to the limitations of classical genetic techniques such as microsatellite-based screens and low density marker sets, except exome sequencing [15]. We here analysed the whole genome sequence of eight samples from one extended MS-enriched family to asses for rare, disease-relevant variants.

3.3 Methods

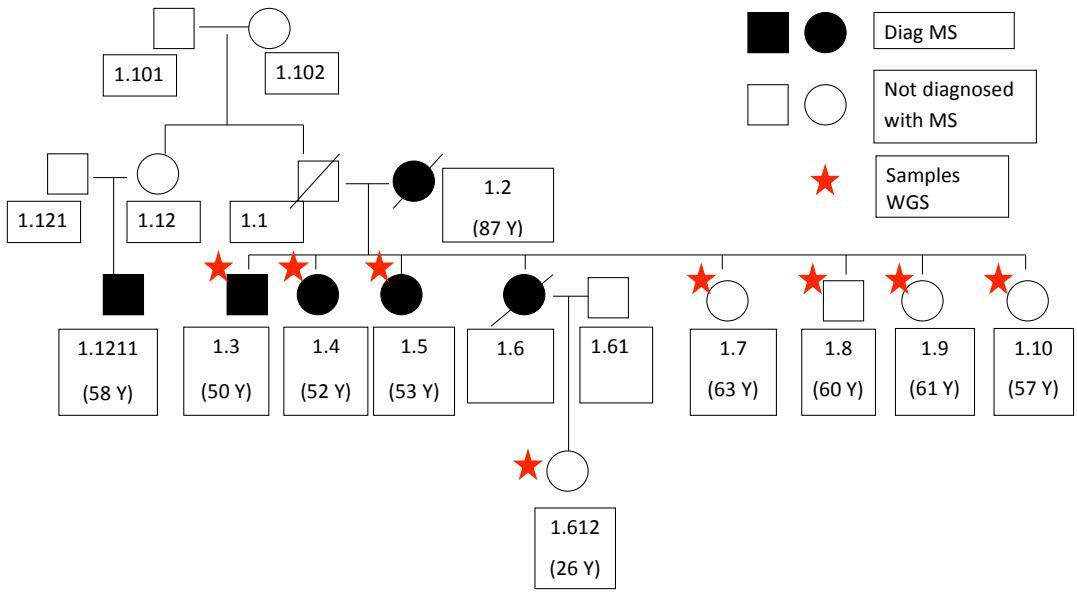
3.3.1 Study subjects

Eight samples from one extended pedigree (shown in **Figure 3.1**), with the proband sourced from Tasmania, Australia, were selected for sequencing, including three MS affected and four unaffected siblings and a daughter (26 years old) of a deceased sibling

with known MS. She however has an unknown disease status due to her age (**Figure 3.1**). Blood or saliva samples were collected from each of the eight subjects seen at the Royal Hobart Hospital or by their local neurologists in Ireland and Sydney, Australia were collected and stored at -80.C. MS cases were diagnosed and confirmed by the study neurologist based on the 2010 McDonald diagnosis criteria [22].

Figure 3.1 Pedigree of the MS family.

Individuals who provided sample for WGS are indicated by the star. Age at blood/saliva collection is shown. Three additional family members were diagnosed with MS but were deceased at the time of collection.



3.3.2 Whole genome sequencing

Whole genome sequencing was conducted at the University of Queensland Diamantina

Institute (UQDI). Genomic DNA was extracted from whole blood or saliva, such that gDNA was extracted from whole blood using a variety of standard laboratory approaches; or gDNA was isolated from saliva self-collected into Oragene DNA tubes according to the manufacturer's instructions (DNAgenotek). DNA concentrations were assessed using pico green fluorescence, ultraviolet (OD260nm) spectrophotometry or on an ethidium bromide–stained low-percentage agarose gel compared to a high-molecular-weight standard. Because of possible bacterial gDNA contamination and difficulty in obtaining reliable pico green and spectrophotometry measurements, all saliva DNA samples were estimated and assessed for their integrity by agarose gel electrophoresis and using at least one other method. Wet-lab sequencing included preparation of eight Illumina TruSeqDNA libraries and sequencing run on an Illumina HiSeq system using 2x100bp paired-end reads at an average coverage (read depth) of 30x.

Sequencing libraries were constructed using a modification of the Illumina TruSeqDNA sample preparation kit. Briefly, 1.6ug of genomic DNA was sheared to an average fragment size of 200bp using the Covaris E220. Fragments were purified using AmpPureXP beads (Beckman Coulter) to remove small products (<100bp), yielding 1ug of material which was end-polished, A-tailed and adapter ligated according to the manufacturer's protocol. The libraries were subjected to minimal PCR cycling and quantified using the Agilent High Sensitivity DNA assay. Each library incorporated a unique barcode sequence allowing all eight to be multiplexed together at an equimolar ratio for subsequent sequencing. Libraries were assessed for both quality and yield using the Agilent High Sensitivity DNA assay and KAPA Library Quantification Kit.

WGS was performed with eight multiplexed samples per flow cell lane using the Illumina

HiSeq2000 platform and TruSeqSBS version 3 reagents to generate 100bp paired-end reads (2x100bp). The pool was sequenced across 27 lanes generating greater than 30x coverage per sample (>100gigabases per sample).

3.3.3 Analyses of whole genome sequencing data

The Illumina Data Analysis Pipeline software (CASAVA 1.8.2) was used for initial base calling and multiplexing of data. Sequence data were aligned to the current build of the human genome (hg19, released February 2009) using the Novoalign alignment tool (V2.08.02)[23], sequence alignment files were converted using SAMtools (v0.1.14)[24] and Picard tools (v1.42). SNPs and indels were called using the Genome Analysis Toolkit (GATK v5506)[25] and annotated using ANNOVAR[26].

Further analysis of sequence data was performed using custom scripts employing R and Bioconductor. We retained good quality SNPs and indels (minimum depth of coverage for SNP calling: >10-fold for homozygous SNPs, >15-fold for heterozygous SNPs). Additionally, we used variants that passed GATK Variant Quality Score Recalibration that incorporates quality parameters, including sequencing depth and quality scores at the SNP position, maximal length of the homopolymer run and strand bias). Remaining SNPs and indels were assessed according to prediction of potentially damaging consequence ("nonsynonymous SNV", "splicing", "frameshift substitution", "stopgain SNV", "stoploss SNV") using Ensemble, RefSeq and UCSC transcripts. Further filtering excluded SNPs with a minor allele frequency (MAF)>0.001 observed in NCBI dbSNP (release 135), 1000Genomes [27], 1000Genomes small indels (called using DINDEL [28]), SNPs from

the NHLBI-ESP (6,500 exome sequencing project) and 956 control exomes generated internally by UQDI. Variants not present in any of these databases were considered novel.

Non-coding elements were accessed by including those only with Genomic Evolutionary Rate Profiling (GERP [29]) score >= 2.0. GERP is a measure of conservation across species, such that neutral sites tend to score near zero, whereas constrained sites generally score positively. The maximum genome-wide score (~5.8) only applies to sites that are perfectly conserved across all sequenced mammals, whereas 'no prediction' applies to sites that are aligned to none or few species (mostly repetitive sequences) [30]. In the case of insertions and deletions the maximum GERP sort within the insertion or deletion location was recorded.

3.3.4 Target variant filtration strategy

Basic transmission analyses were performed using the assumption that the causative variant would be shared (heterozygous or homozygous/dominant or recessive) by the three affected siblings (1.3, 1.4, 1.5). A degree of low penetrance was included by allowing for one unaffected sibling to be heterozygous, which was the initial filtration step. Since we hypothesise that the MS susceptibility variant(s) segregating in this pedigree would be rare or private we filtered variants with minor allele frequency (MAF) ≤0.001 and initially focused only on coding variants. Finally, variants were included if they were present in 10 or more reads, and had pass filter scores.

Next we further filtered the candidate variants according to if the variants were predicted to significantly impact the protein product by at least one of the following *in silico*

prediction tools; Sorting Tolerant From Intolerant' (SIFT) [31] and Polymorphism Phenotying v2 (Polyphen2) [32]. SIFT uses a sequence homology based approach to classify amino acid substitutions [33,34], which is based on the evolutionary conservation of the amino acids within protein families. Briefly, highly conserved positions tend to be intolerant to substitution, whereas those with a low degree of conservation tolerate most substitutions. SIFT can predict correctly 69% of the substitutions associated with the disease to affect protein function [33]. Polyphen2 is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein, such that whether a nonsynonymous SNPs (nsSNP) is damaging or benign [35]. Polyphen2 can achieve the rate of true positive predictions of 92% [32].

Final candidates were filtered based on biological relevance and plausibility. Briefly, genes of interest were investigated using online databases NCBI “Gene” (http://www.ncbi.nlm.nih.gov/gene) and also commercial tools IPA (www.ingenuity.com) to determine their biological function, expression location (such as brain) and potential relevance to MS pathological processes including neurological or immune function.

The WGS data for the final set of variants were manually visualised using Integrative Genomics Viewer (IGV) [36,37], to ensure that the surrounding data and read mapping looked clean prior to confirmation sequencing. Finally, 11 candidate variants of interest were identified after the preceding process of whole genome sequencing.

3.3.5 Validation of candidate variants by Sanger sequencing

Candidate variants identified by WGS analysis were confirmed by Sanger sequencing of

the original eight samples in-house. Primers were designed for PCR as shown in **table 3.1**. PCR reactions were performed in 20ul volumes consisting of 10ul of GoTaq® Green (Promega Coproration, USA), 1.6ul of 10uM for both forward and reverse primers, 1.0ul of 20ng/ul DNA and 5.8ul water. Thermal cycling was performed on a Life Technologies Veriti™ thermal cycler. Conditions were as follows: One cycle of 95°C for 3 minutes, 40 cycles of 95°C for 15 seconds, 60°C for 30 seconds followed by 72°C for 30 seconds, with a final extension of one cycle of 72°C for 1 minute.

Table 3.1. Primer design for the validation of eleven putative mutations identified by whole genome sequencing in MS family members

Chr	Position	Forward Primer	Reverse Primer	Product Size (bp)
2	54080702	tttcttctgtggctttggaga	tgtgtgatcattgtgctgtca	354
3	57850299	ggcaacaagagtgaaactcca	gcctttcattggtgaaatcat	351
6	31921581	aaccaaaaggaatggcctgt	catcgcttctgctgttctca	381
6	51930780	gatgagtcagaaggggcaaa	aaaaattcaactgttgctgga	352
7	122342308	ggtgtaagtctccagacgcaat	atgcctgtcatcctttgacc	394
7	140174293	caggtggatcatgaggtcag	tggaaactacatagtggttgtgaga	387
8	11723096	agctgagtgtggaggcacat	ccaggctggttttcgaact	376
8	144642151	ggcaggagcttccacttcta	cagtggctcacccctgtaat	377
8	408372	aggaacagcctggttctgtc	cctgcttttcagcaaaatgg	411
12	53170912	cccaccaaatacaccaggac	cagacaagtttgcaagaaatcc	403
12	57323282	ctgccccaggaacttacaga	cagttttggaaagggctgag	361

PCR amplicons were purified using Agencourt AMPure XP (Beckman-Coulter) following

the manufacturer's directions. Sequencing reactions were performed in both directions using the BigDye® Terminator v3.1 Cycle Sequencing kit (Life Technologies). Sequencing reactions were subsequently purified using Agencourt CleanSEQ (Beckman-Coulter) following the manufacturer's directions, and sequenced on the ABI Prism® 310 Genetic Analyzer (Applied Biosystems).

Sequence analysis was performed using Sequencher V4.10.1 and variants were assessed by visual inspection. Variants that failed Sanger sequencing were considered artifacts of the next generation sequencing methodology.

3.4 Results

3.4.1 Candidate variants identified by the target variant filtration strategy

The initial filtration step identified 237 variants, which were dominant in the three affected siblings (1.3, 1.4, 1.5) and only one of the unaffected sibling (allowing for incomplete penetrance). Of these, 130 variants had a MAF less than 0.001, and 20 of these variants were identified as coding. Finally, 15 variants passed the GATK quality scoring filters and were considered candidate variants.

When we further assessed whether these above candidate variants were predicted to alter the protein product by using both SIFT[31] and Polyphen2 [32] *in silico* functional prediction algorithms, determined their potential biological plausibility or relevance to MS. We refined our targets to 11 coding variants which were selected for confirmation by Sanger sequencing (**Table 3.2**).

Table 3.2. 11 coding variants were identified from whole genome sequencing analysis

Chr	**Variation**	**Intergenic**	**RefGene**	**Affected siblings hit**	**Unaffected siblings hit**	**PolyPhen2/ SIFT**
2	rs201490058	nonsynonymous SNV	DTNB	3/3	1/4	Unknown/ deleterious
2	rs145914833	nonsynonymous SNV	GPR75	3/3	1/4	Benign/ deleterious
3	rs199942083	nonsynonymous SNV	SLMAP	3/3	1/4	probably_damaging/ tolerated
6	6_31921581	nonsynonymous SNV	RDBP	3/3	1/4	possibly_damaging/ tolerated
6	6_51930780	nonsynonymous SNV	PKHD1	3/3	1/4	Benign/ tolerated
7	7_122342308	nonsynonymous SNV	RNF148	3/3	1/4	Benign/ tolerated
7	rs10709936	intronic	MKRN1	3/3	1/4	*NA/NA
8	8_408372	nonsynonymous SNV	FBXO25	3/3	1/4	Unknown/ deleterious
8	rs71316899	intronic - frameshift	GSDMD	3/3	1/4	*NA/NA
12	12_53170912	nonsynonymous SNV	KRT76	3/3	1/4	Unknown/ deleterious
12	12_57323282	nonsynonymous SNV	SDR9C7	3/3	1/4	Unknown/ deleterious

* Intronic SNPs cannot be predicted by PlolyPhen2 and SIFT software.

3.4.2 Validation of candidate variants confirmed by Sanger sequencing

Eleven candidate variants identified by the WGS analysis were re-sequenced using an alternate sequencing technology (Sanger) to ensure the variants were real and not artifacts of the WGS methodology. One variant (rs201490058 in chromosome 2, refGene: *DTNB*) was not sequenced successfully as this SNP is in a lengthy homopolymer region, another candidate (rs71316899 on chromosome 8, refGene *GSDMD*) failed to replicate in the resequencing as the genotype was not consistent with WGS, leaving nine candidate variants that were confirmed as real **(Table 3.3)**.

Amongst the nine validation candidate variants, six were novel variants and three were known variants. Of the novel variants, chr6_31921581 locates in the gene RD RNA-binding protein *(RDBP)*, which has been demonstrated as associated with MS [8]. In addition, chr6_51930780 locates within the gene of fibrocystin isoform 1 (*PKHD1)*, which has been demonstrated as associated with amyotrophic lateral sclerosis (ALS) [38]. The other four novel variants have been shown to be potentially relevant to neurological disorders, with *RNF148* (chr7_122342308) shown to be involved in the prenatal and postnatal stage of brain development [39], and *FBXO25* (chr8_408372) involved in the postnatal stage of brain development and expressed in the layer1 cortical glutamatergic neurons [39]; or the variants were as transcript factors and associated with tumor (chr12_53170912 in *KRT76* and chr12_57323282 in *SDR9C7*). Of the known variants, two are rare variants (rs145914833 in *GPR75* that express in brain; rs199942083 in *SLMAP* that was associated with type II diabetes) and one is a common variant (rs10709936 in *MKRN1* that associates with two neurological disease: Depressive Disorder [40] and Schizophrenia) (**Table 3.4**). Of these

candidate variants, the two most interesting variants were chr6_31921581 that locates in *RDBP* and chr6_51930780 that locates in *PKHD1*.

Interestingly, when we further tested whether the two interesting coding novel variants (chr6_31921581 and chr6_51930780) associated with MS in the combined MS GWAS dataset comprised 3,543 cases and 5,898 controls [41], we found that in the region of chr6: 319195783- 31929014 that includes the novel variant of chr6_31921581 (*RDBP*), two genome-wide significant signals were identified with a p-value of 1.799e-37 (OR=1.49) and 1.257e-52 (OR=1.63) for rs2072633 (chr6_31919578, upstream about 2kb to chr6_31921581) and for rs437179 (chr6_ 31929014, downstream about 7.4kb to chr6_31921581), respectively. When we further assessed the association in the Tasmanian group that comprised 308 cases and 617 controls, we found that the p-value for rs2072633 and rs437179 was 5.504e-07 (OR=1.65) and 1.202e-06 (OR=1.63), respectively (**Table 3.4**). While no significant association (p>0.05) was found in the region near the novel variants of chr6_51930780 (*PKHD1)* either in the combined GWAS dataset or in the Tasmanian group (**Table 3.5**).

Table 3.3. Candidate genes confirmed by Sanger sequencing

Ref gene	KRT76	SDR9C7	FBXO25	GSDMD	RDBP	PKHD1	RNF148	MKRN1	GPR75	SLMAP
Chr	12	12	8	8	6	6	7	7	2	3
Variant	12_53170912	12_57323282	8_408372	rs71316899	6_31921581	6_51930780	7_122342308	rs10709936	rs145914833	rs199942083
Primer	***C/A***	***G/A***	***G/A***	***-/AGGGCAGGGC***	***G/T***	***T/C***	***G/A***	***A/-***	***A/C***	***A/T***
*BTCK-1.3	C/A	G/A	G/A	&I/D	G/T	T/C	G/A	A/-	A/C	A/T
*BTCK-1.4	C/A	G/A	G/A	I/I	G/T	T/C	G/A	A/-	A/C	A/T
*BTCK-1.5	C/A	G/A	G/A	I/D	G/T	T/C	G/A	A/-	A/C	A/T
#BTCK-1.7	C/A	G/A	G/G	I/D	G/G	T/T	G/G	A/A	A/C	T/T
#BTCK-1.8	C/C	G/G	G/A	I/D	G/T	T/C	G/A	A/-	A/A	A/T
#BTCK-1.9	C/C	G/G	G/G	I/I	G/G	T/T	G/G	A/A	A/A	T/T
#BTCK-1.10	C/C	G/G	G/G	I/I	G/G	T/T	G/G	A/A	A/A	T/T
@BTCK-1.612	C/C	G/G	G/G	I/D	G/G	T/T	G/G	A/A	A/A	T/T
Note	Real	Real	Real	Fail	Real	Real	Real	Real	Real	Real

& I/D refers to insertion/deletion; * affected siblings; # unaffected siblings; @ unknown niece of the seven siblings.

Table 3.4. Description of candidate genes

Chr	Variant	MAF (%)*	RefGene	Gene function&	Associated with disease
2	rs145914833	0.04	GPR75	Cell surface receptor; highly expressed in brain and spinal cord	NA
3	rs199942083	0.1	SLMAP	Sarcolemma associated protein	Type 2 diabetes [42,43]
6	6_31921581	NA	RDBP	Represses RNA polymerase II transcript elongation	MS [8]
6	6_51930780	NA	PKHD1	Correct bipolar cell division; a receptor protein	ALS [38]
7	7_122342308	NA	RNF148	Calcium-binding protein; regulates neurotrophin; mediates Ca(2+)-dependent exocytosis	NA
7	rs10709936	50	MKRN1	Affect on RNA polymerase II-dependent transcription.	Depressive Disorder [40]; Schizophrenia
8	8_408372	NA	FBXO25	Destructing cardiac specific transcription factors. Encoding a brain-expressed F-box protein.	NA
12	12_53170912	NA	KRT76	Responsible for the structural integrity of epithelial cells	Oral carcinogenesis [44]
12	12_57323282	NA	SDR9C7	Similar with the short-chain dehydrogenase/reductase (SDR) family	Oesophageal carcinoma [45]

* The MAF of those known SNPs obtained from UCSC; for the novel SNPs had no known MAF.
& Gene function based on UCSC and the original papers.

Table 3.5. The association with MS in the interesting regions near chr6_31921581 and chr6_51930780

In the combine GWAS data								
Chr	**SNP**	**bp**	**A1**	**F_A**	**F_U**	**A2**	**P-value**	**OR**
6	rs2072633	31919578	A	0.5566	0.4573	G	1.799e-37	1.49
6	rs4151672	31919830	A	0.03967	0.04655	G	0.03078	0.846

6	rs438999	31928306	G	0.07782	0.09109	A	0.00234	0.842
6	rs437179	31929014	A	0.4282	0.3147	C	1.257e-52	1.631
6	rs2082105	51919253	A	0.1227	0.1238	G	0.8353	0.9902
6	rs4715273	51929476	A	0.4837	0.4764	C	0.3478	1.03
6	rs1896981	51936258	G	0.313	0.3072	A	0.4199	1.027
6	rs10484883	51941936	A	0.2442	0.2368	G	0.2648	1.041
In the Tasmanian GWAS data								
Chr	**SNP**	**bp**	**A1**	**F_A**	**F_U**	**A2**	**P-value**	**OR**
6	rs2072633	31919578	A	0.5146	0.3912	G	5.504e-07	1.65
6	rs4151672	31919830	A	0.0276	0.03809	G	0.2446	0.7168
6	rs438999	31928306	G	0.09416	0.09319	A	0.9465	1.011
6	rs437179	31929014	A	0.4643	0.3476	C	1.202e-06	1.626
6	rs2082105	51919253	A	0.1169	0.1159	G	0.9496	1.01
6	rs4715273	51929476	A	0.4886	0.47	C	0.4498	1.077
6	rs1896981	51936258	G	0.3084	0.3015	A	0.7582	1.033
6	rs10484883	51941936	A	0.2419	0.2253	G	0.4245	1.097

3.5 Discussion

In this study, we conducted whole genome sequencing on a unique MS family, for its numbers of affected individuals in one generation. We particularly focused on the siblings (three affected and four unaffected), to find candidate variants that segregated with disease. The family appeared to be segregating MS in an autosomal dominant manner, which suggests that this family may harbor dominant mutations, and therefore re-sequencing using a different platform is an appropriate strategy. Initially, we identified 11 coding variants by whole genome sequencing, where all the affected siblings carried the disease-relevant variants and low penetrance where one unaffected sibling carried one of the disease-relevant variants. The filtering process was delicate: the challenge is to reduce the number of variants to a manageable number of candidates and yet not to lose a possible causative variant in the process. Next, the 11 candidate variants were further validated by Sanger sequencing, and nine candidate variants were identified to be potential disease-causing variants of MS, including six novel variants and three known variants.

Of the nine candidate variants, we are particularly interested in two coding novel variants, chr6_31921581 and chr6_51930780. The variant of chr6_31921581 locates in the gene of *RDBP*, which is part of a complex termed negative elongation factor (NELF) that represses RNA polymerase II transcript elongation. *RDBP* localises to the major histocompatibility complex (MHC) class III region on chromosome 6, which has been shown to be implicated in MS as a risk allele (OR=1.54, p=1.51E-10 for rs2072633) in MS GWAS [8]. Furthermore, the observation of a GERP score of +5.20 for chr6_31921581 was also in keeping with a deleterious genetic change [30], and Polyphen2 and SIFT also showed evidence of an impact on protein function.

Interestingly, further analysis in our combined MS GWAS dataset (described in chapter 2) also supports that the variant locates to a region that is genome-wide significantly associated with MS, and significantly associated with MS in the Tasmanian group. Similar to the genes *CYP27B1* [15] and *TYK2* [46] where both rare functional variants and common variants were identified as associated with MS [13], we detected a novel functional variant in *RDBP* that also contributes to the disease independently of the common variants. As for *PKHD1*, which is predicted to have a single transmembrane (TM)-spanning domain and multiple copies of an immunoglobulin-like plexin transcription-factor domain. *PKHD1* may be required for correct bipolar cell division through the regulation of centrosome duplication and mitotic spindle assembly, and may be a receptor protein that acts in collecting-duct and biliary differentiation. Although no previous evidence showed this gene was associated with MS, we found in this study that the variant of chr6_51930780 was dominant in the affected siblings, and Polyphen2 and SIFT also support chr6_51930780 as having a functional impact on protein structure. Previous, research has linked *PKHD1* with ALS [38], a neurological disease with some similarities to progressive MS, suggesting that chr6_51930780 may be a disease-causing variant in MS. Although no clear genetic overlap between MS and ALS has been detected [47].

With the advent of next-generation sequencing (NGS) technologies, the discovery of novel disease genes is expected to accelerate [48]. Exome sequencing studies in MS families have successfully identified rare variants in the genes *CYP27B1* [15] and *TYK2* [46]. These successful studies showed that small, pedigree based sequencing studies are much more powerful for rare disease gene identification than large scale population based GWASs. Additionally, due to the high cost and the challenge of filtering and interpreting

the large number of identified variants, family-based whole-genome sequencing study is a powerful alternation to population-based whole genome sequencing, in spites of the fact that family-based whole genome analysis can be effective for finding candidate genes as sequence accuracy is enhanced. In our study, we filtered the candidate variants to coding ones, which is similar to exome sequencing, and our findings support that family-based whole genome sequencing –filtering to coding variants is an effective way to detect disease-causing variants.

This chapter presents just our preliminary findings. Next we will genotype the two most interesting candidate variants in a wider cohort of MS cases and controls. In addition, we plan to re-analyse the raw WGS data using our collaborators' pedigree-based variant detection pipeline at the Texas Biomedical Research Institute, San Antonio, USA. This pipeline utilises kinship information to increase the confidence of calls for rare and private variants. We will also extend our analyses to regulatory and non-coding variants following the pedigree-based variant re-call.

Although we could not definitively demonstrate that the most promising disease-relevant variants represent a pathogenic mutation based on our preliminary findings, our findings have demonstrated several potential candidate disease-causing variants to MS, which will build up the foundation for further research, such as validation with our exome array data. Meanwhile, based on the current work, replication will be conducted in another 2 multiplex MS families in Tasmania, with DNA collected from three affected siblings, the parents and on affected maternal aunt in one family and 3 affected siblings 3 unaffected siblings and an unaffected mother in the other. Our preliminary findings, we believe as a useful clue, can open a new avenue for further genetic research on MS.

3.6 Conclusion

We have applied whole genome sequencing on a MS family to detect rare/disease-causing variants, and nine candidate variants were identified with further validation by Sanger sequencing. Two novel coding variants located in *RDBP* and *PKHD1 are* biologically plausible promising potential disease-causing variants to MS, which will be validated in further research. Our study showed a reliable strategy for detecting rare/disease-causing variants and our preliminary findings suggest a foundation for further MS genetic research and replication.

3.7 Postscript

The chapter has described how to detect disease-causing variants of MS by using family-based whole genome sequencing analysis, and we presents our preliminary findings: nine potential disease-relevant candidate variants were identified, where two most interesting candidate variants locate in *RDBP* and *PKHD1* were further tested in a large cohort of MS cases and controls. In the next chapter I will discuss how to detect another component of missing heritability of MS – gene-environment interactions. I will specifically focus on gene-vitamin D interaction, to assess whether gene-vitamin D interactions influencing MS clinical course via a well-validated prospective cohort study design.

3.8 References

1. Sadovnick AD, Armstrong H, Rice GP, Bulman D, Hashimoto L, et al. (1993) A population-based study of multiple sclerosis in twins: update. Ann Neurol 33: 281-285.
2. Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, et al. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 11: 504-510.
3. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
4. Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, et al. (2009) Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet 17: 1309-1313.
5. Ban M, McCauley JL, Zuvich R, Baker A, Bergamaschi L, et al. (2010) A non-synonymous SNP within membrane metalloendopeptidase-like 1 (MMEL1) is associated with multiple sclerosis. Genes Immun 11: 660-664.
6. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767-778.
7. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet 41: 776-782.
8. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851-862.
9. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2009) CD226 Gly307Ser association with multiple autoimmune diseases. Genes Immun 10: 5-10.
10. International Multiple Sclerosis Genetics Consortium (IMSGC). (2010) Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. Hum Mol Genet 19: 953-962.
11. Jakkula E, Leppa V, Sulonen AM, Varilo T, Kallio S, et al. (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. Am J Hum Genet 86: 285-291.
12. Sanna S, Pitzalis M, Zoledziewska M, Zara I, Sidore C, et al. (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 42: 495-497.
13. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
14. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet.

15. Ramagopalan SV, Dyment DA, Cader MZ, Morrison KM, Disanto G, et al. (2011) Rare variants in the CYP27B1 gene are associated with multiple sclerosis. Ann Neurol 70: 881-886.
16. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med 63: 35-61.
17. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.
18. Eraksoy M, Kurtuncu M, Akman-Demir G, Kilinc M, Gedizlioglu M, et al. (2003) A whole genome screen for linkage in Turkish multiple sclerosis. J Neuroimmunol 143: 17-24.
19. Hensiek AE, Roxburgh R, Smilie B, Coraddu F, Akesson E, et al. (2003) Updated results of the United Kingdom linkage-based genome screen in multiple sclerosis. J Neuroimmunol 143: 25-30.
20. Kenealy SJ, Babron MC, Bradford Y, Schnetz-Boutaud N, Haines JL, et al. (2004) A second-generation genomic screen for multiple sclerosis. Am J Hum Genet 75: 1070-1078.
21. Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, et al. (2005) A high-density screen for linkage in multiple sclerosis. Am J Hum Genet 77: 454-467.
22. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, et al. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50: 121-127.
23. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11: 473-483.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
25. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
26. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
27. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, et al. (2012) The 1000 Genomes Project: data management and community access. Nature Methods 9: 459-462.
28. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, et al. (2011) Dindel: accurate indel calls from short-read data. Genome Res 21: 961-973.
29. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-913.
30. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628-640.
31. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073-1081.
32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nature Methods 7: 248-249.
33. Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24: 2397-2398.

34. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22: 2729-2734.
35. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591-597.
36. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.
37. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.
38. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, et al. (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 6: 322-328.
39. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. Nature 478: 483-489.
40. Shyn SI, Shi J, Kraft JB, Potash JB, Knowles JA, et al. (2011) Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. Mol Psychiatry 16: 202-215.
41. Lin R, Charlesworth J, Stankovich J, Perreau VM, Brown MA, et al. (2013) Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis. PLoS One 8: e56379.
42. Ding H, Howarth AG, Pannirselvam M, Anderson TJ, Severson DL, et al. (2005) Endothelial dysfunction in Type 2 diabetes correlates with deregulated expression of the tail-anchored membrane protein SLMAP. Am J Physiol Heart Circ Physiol 289: H206-211.
43. Chen X, Ding H (2011) Increased expression of the tail-anchored membrane protein SLMAP in adipose tissue from type 2 Tally Ho diabetic mice. Exp Diabetes Res 2011: 421982.
44. Ambatipudi S, Bhosale PG, Heath E, Pandey M, Kumar G, et al. (2013) Downregulation of Keratin 76 Expression during Oral Carcinogenesis of Human, Hamster and Mouse. PLoS One 8: e70688.
45. Tang S, Gao L, Bi Q, Xu G, Wang S, et al. (2013) SDR9C7 promotes lymph node metastases in patients with esophageal squamous cell carcinoma. PLoS One 8: e52184.
46. Dyment DA, Cader MZ, Chao MJ, Lincoln MR, Morrison KM, et al. (2012) Exome sequencing identifies a novel multiple sclerosis susceptibility variant in the TYK2 gene. Neurology 79: 406-411.
47. Goris A, van Setten J, Diekstra F, Ripke S, Patsopoulos NA, et al. (2013) No evidence for shared genetic basis of common variants in multiple sclerosis and amyotrophic lateral sclerosis. Hum Mol Genet.
48. Tsuji S (2010) Genetics of neurodegenerative diseases: insights from high-throughput resequencing. Hum Mol Genet 19: R65-70.

Chapter 4. Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis

4.1 Preface

The previous two chapters (chapter 2 and chapter 3) described how to detect one component of missing heritability of MS- rare genetic variants and/or disease-causing variants, by using identity-by descent (IBD) mapping and family-based whole genome sequencing analysis. This chapter will describe how to detect another component of missing heritability of MS – gene-environment interactions. Among the well-estimated environmental factors that were associated with MS onset, low ultraviolet radiation (UVR) exposure and low vitamin D levels are among the strongest and most consistent. In this chapter, I will specifically focus on gene-vitamin D interaction, to assess whether gene-vitamin D interactions influencing MS clinical course via a well-validated prospective cohort study design. Most parts of this chapter have been published in *Journal of Neurology, Neurosurgery & Psychiatry*. 2014; 85(4): 399-404. (Appendix 4A).

4.2 Introduction

Multiple sclerosis is a complex disorder of the centre nervous system (CNS) where gene-environment interactions are considered a key part of disease susceptibility [1]. Among the environmental factors thought to be associated with MS, low UVR exposure and low vitamin D levels are among the strongest and most consistent [2,3]. Studies investigating

vitamin D for its role in MS clinical course have found levels of the major circulating metabolite of vitamin D, 25-hydroxyvitamin D (25(OH)D), are lower during relapse relative to remission [4,5], and inversely associated with relapse risk [6-8]. The inverse relationship between vitamin D and relapse is likely mediated by the immunomodulatory effects of the active metabolite of vitamin D, 1,25-dihydroxyvitamin D ($1,25(OH)_2D$), acting to up-regulate regulatory T-cell function and depress inflammatory immune activity [9].

Genome-wide association studies (GWAS) have successfully identified more than 100 loci as associated with MS risk [10-12], and other studies have suggested a predictive gene-environment interactions between either childhood UVR exposure or vitamin D intake with MS risk and vitamin D receptor (VDR) polymorphisms [13,14]. Less success has been received in the search for genetic modulators of MS clinical course [15] and disease severity [11]. This may be due to the narrow time-frame during which relapses occur and during which genes may exert their effects on relapse risk, directly and via interaction with environmental factors.

We were interested in whether the magnitude of the association between 25(OH)D and relapse was dependent on the genotype. Therefore, using a well-validated prospective cohort study design, we evaluated whether there was interaction between serum 25(OH)D, a number of genetic predictors involved in the vitamin D metabolism and VDR/ retinoid X receptor (RXR) transcription factor formation pathway and subsequent hazard of relapse in MS.

4.3 Materials and methods

4.3.1 Study design

The Southern Tasmanian Multiple Sclerosis Longitudinal Study was designed as a prospective cohort study to evaluate the role of UV exposure and 25(OH)D on the clinical course of MS [6,16]. Briefly, this study followed a cohort of 198 persons with clinically definite MS (2001 McDonald criteria [17]) living in southern Tasmania, Australia between 2002 and 2005. Of these, 145 participants of the relapsing-remitting MS (RRMS) phenotype were followed beyond one review and 141 participants had genotype data. When participants discontinued participation or were lost to follow-up (8/198; 4%), they were censored at the date of study exit or their last attended review. Ethics approval was obtained from the Southern Tasmania Human Research Ethics Committee. All participants provided informed consent.

4.3.2 Measurement of relapses and 25(OH)D

These measurements have been described in detail elsewhere [6]. Briefly, relapses were defined according to established criteria [17] and were reported by phone or at each 6-monthly review. All relapse reports were validated by the study physician and study neurologist.

At each summer and winter review (January-April, and June-September), blood samples were taken. All samples were stored at -80°C and shielded from light. Since 25(OH)D is the major circulation form of vitamin D, provides the best estimate of a patient's long-term vitamin D status [18], and has been associated with MS onset and relapse, free and bound serum 25(OH)D concentrations were measured using a commercially available radioimmunoassay (Stillwater, Minnesota-DiaSorin Inc). Assays were performed

following the conclusion of the study. Consequently, neither participants nor study personnel were aware of participants' 25(OH)D concentrations during the study.

4.3.3 Genotyping

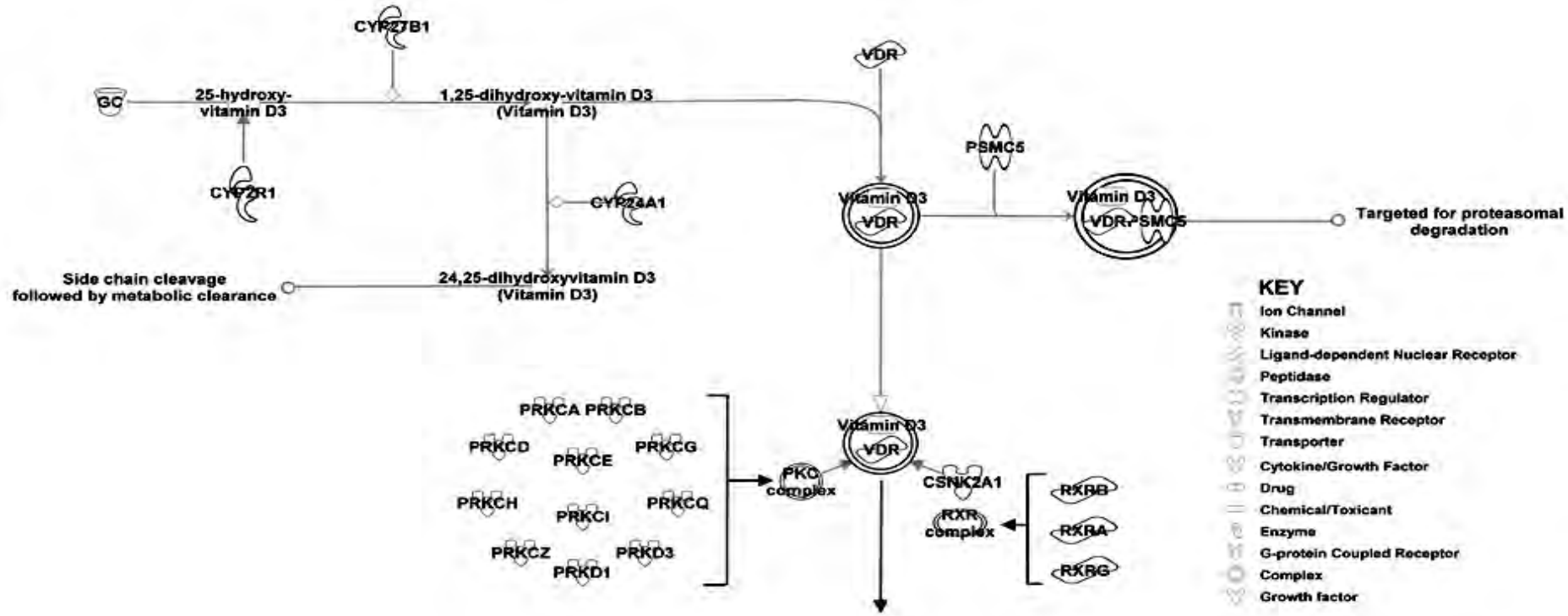
A total of 164 MS cases were genotyped on the Illumina Infinium Hap370CNV Array as a part of ANZgene MS GWAS [19]. An additional 29 MS cases were genotyped using the Illumina HumanOmniExpress-12v1_A array. All the samples were previously identified as being from persons of European descent [19], and a conservative quality control was conducted with PLINK [20]: individuals with call rates less than 0.90, single nucleotide polymorphisms (SNPs) with call rates less than 0.95 or in Hardy-Weinberg equilibrium ($p<10^{-7}$), or duplicates were excluded, leaving 189 cases with 290,536 SNPs.

4.3.4 Vitamin D pathway analysis and SNP selection

It has long been recognised that genes do not work alone, but in an intricate network of interaction. As our *a priori* hypothesis was that there are gene-vitamin D interactions which modulate the clinical course of MS, we therefore generated a canonical vitamin D metabolism and VDR/RXR transcription factor formation pathway using Ingenuity Pathway Analysis (IPA) (www.ingenuity.com), which includes an extensive Knowledge Base derived from published interactions between gene products and the different forms of vitamin D. The pathway of vitamin D metabolic processes and formation of the VDR/RXR transcription factor complex is shown in **Figure 4.1**. In total, there were 21 genes involved in this pathway. For each of these genes we selected all SNPs from the Hg19 genome and 1000 Genome data [21] with minor allele frequency (MAF) ≥1% located within the physical boundaries, including 1kb upstream or downstream, and the

genotyped SNPs for each of these genes were filtered further based on the genotype dataset. Finally, 276 genotyped SNPs from 21 genes were tested (**Supplementary Table 4.1**), for those SNPs with high linkage disequilibrium (LD) in the same gene, TagSNPs were selected using the r^2-based tagger tool SNAP [22].

Figure 4.1. The vitamin D metabolism and VDR/RXR transcription factor complex formation pathway. The vitamin D metabolism and VDR transcription factor complex formation pathway was generated using the canonical pathway information stored in IPA (Ingenuity® Systems, www.ingenuity.com). Genes or gene products are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All relationships are supported by multiple published references. The shape of each node represents the functional class of the gene product, as shown in the key.



Height (m) and weight (kg) were measured at study entry, and body mass index (BMI) was calculated as weight divided by height squared. Also at study entry, skin melanin density was measured on the upper inner arm using a spectrophotometer [23]. Clinical disability was measured every winter review by a single physician, including the Expanded Disability Status Scale (EDSS).

4.3.5 Data analysis

Predictors of 25(OH)D were evaluated by multilevel mixed-effects linear regression (SEs inflated to allow for unknown covariance structure), with SNPs as the predictors and 25(OH)D as the outcome adjusted for season of measure (summer, winter), age, BMI, melanin density (%), current sun exposure ('less than 0.5 h a day'; '1/2 to 1 h a day'; '1 to 2 h a day'; '2 to 3 h a day'; '3 to 4 h a day'; 'greater than 4 h a day'), vitamin D supplementation ('never' ; 'less than 200 IU a day'; '200 to 400 IU a day'), fish intake ('never'; 'less than once a week'; '1 to 2 serves a week'; '2 to 4 serves a week'; 'greater than 4 serves a week') and smoking (No, Yes). Dependent variables were transformed as required to make the residuals less heteroskedastic; however all regression coefficients are reported on the scale of the original variable.

The hazard of relapse was modelled as a function of 25(OH)D and other covariates on time-to-relapse and was calculated using Cox proportional hazards models for repeated events, using the gap-time model described by Prentice and colleagues [24]. Standard errors were adjusted to reflect multiple events per person. All covariates satisfied the proportional hazards assumption with the exception of the binary variable for sex and the categorical variable for baseline EDSS (0 – 2.5, 3.0 – 5.0, 5.5– 7.0, 7.5 – 9.0). For this reason, all models are stratified to allow the baseline hazards to differ by sex and baseline

EDSS category. 25(OH)D was estimated at monthly intervals between summer and winter measurements using methods described previously [6]. Briefly, a sinusoidal curve was fitted to the measured 25(OH)D to yield a sample-average value of 25(OH)D over the year, this used to predict values at monthly intervals between measures for each person. The sinusoidal regression model was:

$$25(OH)D \text{ levels} = \beta_0 + \beta_1 \sin\frac{2\pi t}{365} + \beta_2 \cos\frac{2\pi t}{365}$$

where *t* denotes the day of the year the sample was collected, and $\beta j (j=0,1,2)$ are estimated regression coefficients. These modelled 25(OH)D values were then used as the primary predictor variable in survival analysis models. To examine whether there was an interaction between modelled 25(OH)D and a SNP on relapse, a product term was included in the model, and then the association between 25(OH)D and hazard of relapse can be estimated at each SNP allele level.

To estimate cumulative effects of the significant and other top SNPs involved in the vitamin D pathway on relapse or on 25(OH)D levels, we created a variable that provided values for the number of risk genotypes to represent the "genetic risk score''. Briefly, a person was designed as having a risk variant of a SNP if they carried a SNP that was independently associated with increased hazard of relapse. For instance, a person for whom the heterozygote and minor homozygote were each associated with an increased hazard of relapse, all persons heterozygous or homozygous for the minor allele would be considered carriers of the risk variant, whereas homozygotes for the major allele would be considered non-carriers of the risk variant. The total number of risk variants was summed and the risk score included as a covariate in the model.

Significance threshold was adjusted by Bonferroni correction for the number of genes evaluated. SNPs with a *p*-value < 0.00238 (0.05/21), or the adjusted *p*-value (P_{adj})< 0.05 after adjustment for covariates and multiple comparisons for all 21 genes were considered to be significant.

All data analyses were performed using STATA/IC V.12.1 (StataCorp LP, College Station, Texas, USA).

4.4 Results

The total of participants who were of relapsing-remitting MS (RRMS) course and followed beyond one review, and for whom SNP data were obtained, comprised the cohort for analysis. This group was followed up for an average of 2.3 years, included 75.2% females, and had mean age 45.9 (SD, 10.2) years, and 82.3% (116/141) used immunomodulatory therapy during the study. The mean EDSS at study entry was 2.8 (SD, 1.5), and the mean MS duration from diagnosis was 6.95 year (SD, 7.08). A total of 122 confirmed relapses occurred in 70 participants.

We found that the relationship between 25(OH)D and hazard of relapse was significantly different for different alleles of two genotyped SNPs (rs908742 in *PRKCZ* and rs3783785 in *PRKCH*), even after adjustment for covariates (age, sex and baseline EDSS) and this interaction persisted after adjustment for multiple comparisons ($p_{interaction}$=0.001, p_{adj}=0.021, respectively; **Table 4.1**). For example, a significant protective association was found between 25(OH)D and relapse when subjects carried the major homozygous genotype of rs3783785 (HR=0.76, p=4.65x10^{-6}) and rs908742 (HR=0.77, p=3.63x10^{-5}).

However, it was not significant when subjects carried at least one copy of the rare allele of rs3783785, and an increased risk was observed when individuals carried the minor homozygous genotype of rs908742 (HR=1.18, *p*=0.017; **Table 4.2).** The modulating effect of the two SNPs also persisted after adjustment for immunomodulatory medication use (data not shown). The two SNPs were not associated with hazard of relapse or levels of 25(OH)D (**Table 4.1**).

Table 4.1. The significant genotyped SNPs and their associations with 25(OH)D levels and the 25(OH)D-relapse association

2 genotyped SNPs modified the relationship between 25(OH)D and the hazard of relapse										
Chr	SNP	bp	Allele	MAF	HR_25(OH)D (95%CI)	*P_interact*	*P_adj_interact*	*P_relap*	*P_25(OH)D*	Nearest Gene
1	rs908742	2033256	A/G	0.283	0.85 (0.77, 0.94)	0.001	0.021	0.39	0.35	*PRKCZ*
14	rs3783785	61937292	G/A	0.270	0.83 (0.75, 0.92)	0.001	0.021	0.14	0.09	*PRKCH*
2 genotyped SNPs associated with the levels of 25(OH)D (nmol/L)										
Chr	SNP	bp	Allele	MAF	Coef_25(OH)D (95%CI)	*P_25(OH)D*	*P_adj_25(OH)D*	*P_interact*	*P_relap*	Nearest Gene
11	rs1993116	14910234	A/G	0.384	-5.14 (-8.27, -2.01)	0.001	0.021	0.58	0.06	*CYP2R1*
16	rs7404928	23888840	G/A	0.249	-6.24 (-9.88, -2.59)	0.001	0.021	0.10	0.95	*PRKCB*

P_relap: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

P_interact: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

P_25(OH)D: adjusted for sun exposure, season, age, BMI, vitamin D supplementation, melanin density, fish intake and smoking , before adjustment for multiple comparisons.

Padj: adjusted for covariates, after adjustment for multiple comparisons for 21 genes involved in the vitamin D metabolism and information complex pathway.

Bp =base pair (hg19); HR=hazard ratio; 25(OH)D= 25-hydroxyvitamin D; CI= confidence interval.

Table 4.2. The association between 25(OH)D and hazard of relapse stratified by allele for the significant genotyped SNPs that modified the 25(OH)D-relapse association

Chr/gene	SNP	Genotype	N	HR_25(OH)D (95%CI)	*p*
1/PRKCZ	rs908742	AA	14	1.18 (1.02, 1.35)	0.017
		AG	52	0.91 (0.81, 0.999)	0.06
		GG	75	0.77 (0.68, 0.87)	$3.63x10^{-5}$
					$p_{interaction}=0.001$
14/PRKCH	rs3783785	AA	78	0.76 (0.67, 0.85)	$4.65x10^{-6}$
		AG+GG	63	0.98 (0.89, 1.08)	0.732
					$p_{interaction}=0.001$

$p_{interaction}$ was assessed with the SNP coded as 0,1,2 and the model was adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

The GG genotype of rs3783785 (n=9) was combined with the AG genotype because of the low frequency.

We also found two genotyped SNPs (rs1993116 in *CYP2R1* and rs7404928 in *PRKCB*) that were significantly associated with lower levels of 25(OH)D; persisting after adjustment for sun exposure and other relevant confounders and after adjustment for multiple comparisons (p_{trend} =0.001, p_{adj}=0.021, respectively; **Table 4.1**). A clear dose-response was observed with the *CYP2R1* SNP rs1993116, compared to those homozygous for the minor allele heterozygotes had levels of 25(OH)D that were significantly lower by 7.1 nmol/L (p=0.005) and major allele homozygotes were significantly lower by 12.4 nmol/L (p=0.004). Dose-response could not be examined reliably for rs7404928 because

of the low numbers in the minor allele group (**Table 4.3**). These two SNPs were not associated with hazard of relapse, and did not modify the association between 25(OH)D and relapse (**Table 4.1**).

We then examined the combined effect on 25(OH)D levels of the two genotyped SNPs associated with low 25(OH)D (rs1993116 in *CYP2R1* and rs7404928 in *PRKCB*) by creating a 'genetic risk score'. We found that 25(OH)D levels were significantly lower by 13.58 nmol/L (p=0.00025) for those individuals carrying two 'risk' genotypes compared to those carrying less than one 'risk' genotype (**Table 4.3**). This cumulative effects of these two SNPs accounted for 3.5% of the variation in 25(OH)D levels with rs1993116 in *CYP2R1* accounting for 2.6% of the variation.

Table 4.3. The allele dosage response and the cumulative effect of the significant genotyped SNPs that were associated with low 25(OH)D

Chr/gene	SNP	Genotype	N	25(OH)D (95%CI) (nmol/L)	*p*
11/CYP2R1	rs1993116	AA	19	54.13 (50.98, 57.27)	ref
		AG	68	-7.10 (-12.08, -2.12)	0.005
		GG	54	-12.41 (-20.89, -3.94)	0.004
					p trend=0.001
16/PRKCB	rs7404928	GG+GA	68	54.67 (51.03, 58.31)	ref
		AA	73	-7.19 (-11.98, -2.40)	0.003
Cumulative effects of the 2 SNPs associated with lower 25(OH)D					
≤1 genotype *			23	62.31 (55.45, 69.17)	ref
2 genotypes			118	-13.58 (-20.84, -6.31)	0.00025

p_{trend}: adjusted for sun exposure, season, age, BMI, vitamin D supplementation, melanin density, fish intake and smoking, before adjustment for multiple comparisons.
The GG genotype of rs7404928 (n=5) was combined with the GA genotype because of the low frequency.
* Those carried 0 genotype (n=1) was combined with that carried 1 genotype.

We did not identify any SNPs that were significantly associated with hazard of relapse after adjusting for covariates and multiple comparisons. However, four genotyped SNPs (rs281508 and rs6740453 in *PRKCE*, rs3733359 in *GC*, and rs3818740 in *RXRA*; **Table 4.4**) were significantly associated with the hazard of relapse after adjustment for covariates but not multiple comparisons. When we examined the combined effect on relapse of these SNPs creating a 'genetic risk score', we found a dose-response effect with increasing number of risk genotypes (p_{trend} =8.86x10^{-6}). For example, compared to subjects carrying less than two risk genotypes, those with three and four risk genotypes had an increased relapse HR of 1.95 (95% CI 1.17 to 5.25) and 4.21 (95% CI 2.27 to 7.81; **Table 4.4**), respectively.

Table 4.4. The allele dosage response and cumulative effect for the genotyped SNPs that were associated with the hazard of relapse

Chr/gene	SNP	Genotype	N	HR_relap (95%CI)	*p*
2/PRKCE	rs281508	AA	14	1.00 (ref)	ref
		AC	63	1.63 (0.78, 3.38)	0.193
		CC	64	2.21 (1.14, 4.29)	0.02
					p_{trend} =0.019
					p_{adj}=0.40
2/PRKCE	rs6740453	GT	121	1.00 (ref)	ref
		TT	20	1.71 (1.01, 2.89)	0.046
					p_{adj}=0.97
4/ GC	rs3733359	AA+AG	22	1.00 (ref)	ref
		GG	119	1.97 (1.06, 3.66)	0.032
					p_{adj}=0.67
9/ RXRA	rs3818740	CC	19	1.00 (ref)	ref
		CT	57	3.35 (1.13, 9.98)	0.03
		TT	65	3.61 (1.22, 10.68)	0.02
					p_{trend} =0.031
					p_{adj}=0.65
Cumulative risk of relapse for the 4 SNPs					
≤ 2 risk genotypes *			41	1.00 (ref)	ref
3 risk genotypes			87	1.95 (1.17, 3.25)	0.01
4 risk genotypes			13	4.21 (2.27, 7.81)	$5.17x10^{-6}$
					$p_{trend}=8.86x10^{-6}$

p_{trend}: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

None carried GG genotype of rs6740453, and the AA genotype of rs3733359 (n=1) was combined with the AG genotype because of the low frequency.
* Those carried 0 risk genotype (n=1) and 1 risk genotype (n=5) was combined with that carried 2 risk genotypes.

4.5 Discussion

In a prospective MS cohort designed to assess the effects of vitamin D and personal UVR exposure on MS clinical course, we have shown that two loci within the genes involved in the vitamin D pathway interact with an environmental factor, serum 25(OH)D, to

influence the clinical course of MS. Two SNPs (rs908742 in *PRKCZ* and rs3783785 in *PRKCH*) were found to significantly modify the association between serum 25(OH)D and hazard of relapse. In line with this finding, we identified a number of other genes involved in the vitamin D pathway that were associated with hazard of relapse, although these associations were not significant after adjustment for multiple comparisons. We also identified two SNPs (rs1993116 in *CYP2R1* and rs7404928 in *PRKCB*) that were significantly associated with 25(OH)D levels.

Previously, we found that higher 25(OH)D levels were associated with lower relapse risk in this same cohort [6]. Interestingly, the inverse association between 25(OH)D and relapse was observed among those with the major homozygous genotypes of rs3783785 and rs908742, both located within introns of the protein kinase C (PKC) family genes *PRKCH* and *PRKCZ.* PKC family genes have been shown to be associated with other neurological disorders including Alzheimer's disease, status epilepticus and cerebellar ataxia [25]. PKC family genes have been shown to be regulated by $1,25(OH)_2D$ in chondrocytes [26] and mediated by VDR in downstream signalling pathways [27]. Our findings suggest that *PRKCH* and *PRKCZ* do not have strictly independent roles but work in conjunction with each other, as we found that other genes in the PKC family were associated with 25(OH)D levels (*PRKCB*), and hazard of relapse (*PRKCE*), although some of these associations were not significant after adjustment for multiple comparisons. For those associated with relapse, we demonstrated a cumulative effect with increasing number of risk genotypes, with those with three and four risk genotypes having hazard of relapse of 1.95 (95% CI 1.17 to 5.25) and 4.21 (95%CI 2.27 to 7.81; $p_{trend}=8.86x10^{-6}$), respectively.

MS is believed to be the result of a misdirected immune response by autoreactive T-cells against as yet undetermined CNS antigens [28]. *PRKCH*, *PRKCZ, PRKCB* as well as other genes in the 11-member PKC family have been found to affect T-cell activation [29-31]. It is possible therefore that abnormalities in PKC activity induced by one or more of these polymorphisms may lead to altered T-cell function. *PRKCA* has been reported to be associated with MS risk in UK, Finish and Canadian populations [32,33], and *PRKCB* and *PRKCH* had been associated with the risk of similar autoimmune diseases including systemic lupus erythematosus (SLE) [34] and rheumatoid arthritis [35,36]. In rheumatoid arthritis, *PRKCH* messenger RNA was expressed at high levels in T-cells and was significantly down-regulated during immune response [35].

We also identified two SNPs within the *CYP2R1* and *PRKCB* genes that were inversely associated with levels of 25(OH)D. A *CYP2R1* SNP (rs10741657) has previously been associated with vitamin D levels in a vitamin D GWAS [37]. Notably, the SNP (rs1993116) identified in this cohort is in complete LD with rs10741657 (LD=1), which supports the association identified in our cohort. Interestingly, the SNP (rs7404928) in *PRKCB* was suggested to be associated with risk of rheumatoid arthritis [35,36] and was identified in this cohort as a novel polymorphism associated with low 25(OH)D levels, but was not identified in the vitamin D GWAS. *PRKCB* is a member of the PKC gene family that has been demonstrated to be regulated by $1,25(OH)_2D$ in chondrocytes [26] and mediated by VDR in downstream signalling pathways [27]. Studies have shown that *PRKCB* increases phosphorylation of VDR [38] and is a key element in normal T-cell migration [30], and 25(OH)D has been demonstrated to modulate T_{Reg} (regulatory T cell) and T_H cell function in vivo [39], suggesting that *PRKCB* may mediate some of the effect

of 25(OH)D on T cell function. Notably, the cumulative effects of these two SNPs accounted for 3.5% of the variation in 25(OH)D concentrations, in which rs1993116 in *CYP2R1* accounted for 2.6%, which was similar to that identified from the vitamin D GWAS [37].

We now also consider most common diseases to be genetically influenced by the combined effects of many loci and/or a number of rare variants that interact with environmental factors in multiple ways [40]. In this study, we observed a cumulative effect of several SNPs on either the risk of relapse or 25(OH)D levels, supporting the theory that pathway analysis may increase the overall effect within known signalling pathways and enhance the genetic susceptibility to disease. Such an approach has proved useful in examining the contribution of T-cell function, cell adhesion, and cellular communication in the pathogenesis of MS [11,41]. In our study, using pathway analyses, we have been able to demonstrate that several genes potentially work together to influence 25(OH)D levels and/or the hazard of relapse. Consequently, modulation of these genes and/or pathways is an important potential avenue of investigation in the treatment of MS and other immunological disorders.

A strength of our study was the availability of detailed environmental data, including potential confounders such as sun exposure and vitamin D supplementation. A weakness of our study is the sample size. Even though this is one of the largest and most well-studied MS cohorts available, the difficulty of reaching statistical significance reflects the challenge of undertaking genetic studies of clinical course, in comparison with aetiology studies. These questions of gene-environment interactions on MS clinical course can only be answered in longitudinal studies of this nature. Therefore, as discussed above, we have

used other methodologies to overcome potential type 1 error, including allele dose responses and cumulative genotype risk scores. While some of the associations identified here are not significant after adjustment for multiple comparisons, we believe that these results still provide important evidence for real associations. Our results, demonstrate one methodology that can bridge the gap between the need for larger and larger genetic studies and the realities of undertaking MS longitudinal studies.

4.6 Summary

Background: The interplay between genes and environmental factors on MS clinical course has been little studied.

Methods: We conducted a prospective cohort study of 141 participants with relapsing-remitting MS and genotype data followed from 2002 to 2005, and examined genes in the vitamin D metabolism and VDR/RXR transcription factor formation pathway. Gene-vitamin D interactions and the genetic predictors of relapse were assessed using survival analysis. Genetic predictors of 25(OH)D were evaluated by multilevel mixed-effects linear regression. Significance threshold was adjusted by Bonferroni correction for the number of genes evaluated.

Results: The relationship between 25(OH)D and hazard of relapse was significantly different for different alleles of two intronic SNPs (rs908742 in *PRKCZ* and rs3783785 in *PRKCH*) in the protein kinase C (PKC) family genes ($p_{interaction}$=0.001, P_{adj} =0.021 for both). Two other intronic SNPs (rs1993116 in *CYP2R1*and rs7404928 in *PRKCB*) were significantly associated with lower levels of 25(OH)D ($p_{interaction}$=0.001, P_{adj}=0.021 for both). A cumulative effect of multiple 'risk' genotypes on 25(OH)D levels and hazard of

relapse was observed for the significant SNPs (p_{trend}= $7.12x10^{-6}$ for 25(OH)D levels, p_{trend}=$8.86x10^{-6}$ for hazard of relapse).

Conclusion: Our data supports the hypothesis that gene-vitamin D interactions may influence MS clinical course, and that the PKC family genes may play a role in the pathogenesis of MS relapse through modulating the association between 25(OH)D and relapse.

4.7 Postscript

This chapter has discussed how to detect gene-vitamin D interactions in patients with MS in a prospective cohort study. We have demonstrated that PKC family genes modulated the relationship between serum vitamin D and relapse in patients with MS. These findings provided support for our hypothesis that gene-environment interactions influence clinical course of MS. These findings also lead a question to us: did known MS-associated susceptibility interact with serum vitamin D influencing clinical course of MS? Next chapter I will focus on known MS-associated susceptibility. I will examine whether they modulate the relationship between serum vitamin D and relapse by using the same prospective cohort design.

4.8 References

1. Lin R, Charlesworth J, van der Mei I, Taylor BV (2012) The genetics of multiple sclerosis. Practical Neurology 12: 279-288.
2. Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A (2006) Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. Jama 296: 2832-2838.

3. Lucas RM, Ponsonby AL, Dear K, Valery PC, Pender MP, et al. (2011) Sun exposure and vitamin D are independent risk factors for CNS demyelination. Neurology 76: 540-548.
4. Soilu-Hanninen M, Laaksonen M, Laitinen I, Eralinna JP, Lilius EM, et al. (2008) A longitudinal study of serum 25-hydroxyvitamin D and intact parathyroid hormone levels indicate the importance of vitamin D and calcium homeostasis regulation in multiple sclerosis. J Neurol Neurosurg Psychiatry 79: 152-157.
5. Soilu-Hanninen M, Airas L, Mononen I, Heikkila A, Viljanen M, et al. (2005) 25-Hydroxyvitamin D levels in serum at the onset of multiple sclerosis. Mult Scler 11: 266-271.
6. Simpson S, Jr., Taylor B, Blizzard L, Ponsonby AL, Pittas F, et al. (2010) Higher 25-hydroxyvitamin D is associated with lower relapse risk in multiple sclerosis. Ann Neurol 68: 193-203.
7. Runia TF, Hop WC, de Rijke YB, Buljevac D, Hintzen RQ (2012) Lower serum vitamin D levels are associated with a higher relapse risk in multiple sclerosis. Neurology 79: 261-266.
8. Mowry EM, Krupp LB, Milazzo M, Chabas D, Strober JB, et al. (2010) Vitamin D status is associated with relapse rate in pediatric-onset multiple sclerosis. Ann Neurol 67: 618-624.
9. Ascherio A, Munger KL, Simon KC (2010) Vitamin D and multiple sclerosis. Lancet Neurology 9: 599-612.
10. Patsopoulos NA, de Bakker PI, Esposito F, Reischl J, Lehr S, et al. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol 70: 897-912.
11. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
12. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. (2013) Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet.
13. Simon KC, Munger KL, Xing Y, Ascherio A (2010) Polymorphisms in vitamin D metabolism related genes and risk of multiple sclerosis. Mult Scler 16: 133-138.
14. Dickinson JL, Perera DI, van der Mei AF, Ponsonby AL, Polanowski AM, et al. (2009) Past environmental sun exposure and risk of multiple sclerosis: a role for the Cdx-2 Vitamin D receptor variant in this interaction. Mult Scler 15: 563-570.
15. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Human molecular genetics 18: 767-778.
16. Pittas F, Ponsonby AL, van der Mei IA, Taylor BV, Blizzard L, et al. (2009) Smoking is associated with progressive disease course and increased progression in clinical disability in a prospective cohort of people with multiple sclerosis. J Neurol 256: 577-585.
17. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, et al. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50: 121-127.
18. Zerwekh JE (2008) Blood biomarkers of vitamin D status. Am J Clin Nutr 87: 1087S-1091S.

19. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.
21. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, et al. (2012) The 1000 Genomes Project: data management and community access. Nature Methods 9: 459-462.
22. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938-2939.
23. van der Mei IA, Ponsonby AL, Dwyer T, Blizzard L, Simmons R, et al. (2003) Past exposure to sun, skin phenotype, and risk of multiple sclerosis: case-control study. Bmj 327: 316.
24. Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68: 373-379.
25. Kodavanti PR (2011) Cell signaling and neurotoxicity: protein kinase C in vitro and in vivo. Methods Mol Biol 758: 307-319.
26. Schwartz Z, Sylvia VL, Luna MH, DeVeau P, Whetstone R, et al. (2001) The effect of 24R,25-(OH)(2)D(3) on protein kinase C activity in chondrocytes is mediated by phospholipase D whereas the effect of 1alpha,25-(OH)(2)D(3) is mediated by phospholipase C. Steroids 66: 683-694.
27. Boyan BD, Chen J, Schwartz Z (2012) Mechanism of Pdia3-dependent 1alpha,25-dihydroxy vitamin D3 signaling in musculoskeletal cells. Steroids 77: 892-896.
28. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG (2000) Multiple sclerosis. N Engl J Med 343: 938–952.
29. Baier G (2003) The PKC gene module: molecular biosystematics to resolve its T cell functions. Immunol Rev 192: 64-79.
30. Volkov Y, Long A, McGrath S, Ni Eidhin D, Kelleher D (2001) Crucial importance of PKC-□(I) in LFA-1-mediated locomotion of acti- vated T cells. Nat Immunol 2: 508-514.
31. Gomez J, Pitton C, Garcia A, Martinez de Aragon A, Silva A, et al. (1995) The zeta isoform of protein kinase C controls interleukin-2-mediated proliferation in a murine T cell line: evidence for an additional role of protein kinase C epsilon and beta. Exp Cell Res 218: 105-113.
32. Barton A, Woolmore JA, Ward D, Eyre S, Hinks A, et al. (2004) Association of protein kinase C alpha (PRKCA) gene with multiple sclerosis in a UK population. Brain 127: 1717-1722.
33. Saarela J, Kallio SP, Chen D, Montpetit A, Jokiaho A, et al. (2006) PRKCA and multiple sclerosis: association in two independent populations. PLoS Genet 2: e42.
34. Sheng YJ, Gao JP, Li J, Han JW, Xu Q, et al. (2011) Follow-up study identifies two novel susceptibility loci PRKCB and 8p11.21 for systemic lupus erythematosus. Rheumatology (Oxford) 50: 682-688.
35. Takata Y, Hamada D, Miyatake K, Nakano S, Shinomiya F, et al. (2007) Genetic association between the PRKCH gene encoding protein kinase Ceta isozyme and rheumatoid arthritis in the Japanese population. Arthritis Rheum 56: 30-42.
36. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet 44: 511-516.

37. Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, et al. (2010) Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet 376: 180-188.
38. Hsieh JC, Jurutka PW, Nakajima S, Galligan MA, Haussler CA, et al. (1993) Phosphorylation of the human vitamin D receptor by protein kinase C. Biochemical and functional evaluation of the serine 51 recognition site. J Biol Chem 268: 15118-15126.
39. Smolders J, Thewissen M, Peelen E, Menheere P, Tervaert JW, et al. (2009) Vitamin D status is positively correlated with regulatory T cell function in patients with multiple sclerosis. PLoS One 4: e6635.
40. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics 40: 695-701.
41. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18: 2078-2090.

Supplementary table 4.1. Tested genotyped SNPs

No	SNP	Genes nearby/within
1	rs7935792	CYR2R1
2	rs1993116	CYR2R1
3	rs11023374	CYR2R1
4	rs8176345	CYP27B1
5	rs1555439	CYP24A1
6	rs912505	CYP24A1
7	rs927650	CYP24A1
8	rs4809960	CYP24A1
9	rs6097797	CYP24A1
10	rs6097801	CYP24A1
11	rs1352844	GC
12	rs7041	GC
13	rs222020	GC
14	rs705117	GC
15	rs1491709	GC
16	rs3733359	GC
17	rs2665833	PSMC5
18	rs2239186	VDR
19	rs886441	VDR
20	rs2238136	VDR
21	rs2189480	VDR
22	rs1540339	VDR
23	rs1544410	VDR
24	rs4760648	VDR
25	rs2239182	VDR
26	rs2239179	VDR

27	rs4516035	VDR
28	rs2254210	VDR
29	rs2107301	VDR
30	rs205908	CSNK2A1
31	rs4815707	CSNK2A1
32	rs461944	CSNK2A1
33	rs6107514	CSNK2A1
34	rs1010546	PRKCA
35	rs4417581	PRKCA
36	rs12232511	PRKCA
37	rs4536508	PRKCA
38	rs16960110	PRKCA
39	rs759115	PRKCA
40	rs1806448	PRKCA
41	rs8078231	PRKCA
42	rs3889392	PRKCA
43	rs7342847	PRKCA
44	rs6504441	PRKCA
45	rs3889237	PRKCA
46	rs2024321	PRKCA
47	rs7220007	PRKCA
48	rs4644888	PRKCA
49	rs1006841	PRKCA
50	rs877447	PRKCA
51	rs8081512	PRKCA
52	rs7218480	PRKCA
53	rs7220127	PRKCA
54	rs7219495	PRKCA
55	rs8067877	PRKCA
56	rs887797	PRKCA
57	rs8069926	PRKCA
58	rs7207345	PRKCA
59	rs11656099	PRKCA
60	rs9896134	PRKCA
61	rs4790904	PRKCA
62	rs16960114	PRKCA
63	rs8069696	PRKCA
64	rs10491204	PRKCA
65	rs9303510	PRKCA
66	rs16960070	PRKCA
67	rs9908814	PRKCA
68	rs8074995	PRKCA
69	rs4411531	PRKCA
70	rs8464	PRKCA
71	rs956952	PRKCA
72	rs7221968	PRKCA

73	rs9893560	PRKCA
74	rs741141	PRKCA
75	rs3848423	PRKCA
76	rs17753569	PRKCB
77	rs120909	PRKCB
78	rs9922316	PRKCB
79	rs4787676	PRKCB
80	rs8063132	PRKCB
81	rs9925126	PRKCB
82	rs198207	PRKCB
83	rs403018	PRKCB
84	rs2283540	PRKCB
85	rs4788426	PRKCB
86	rs9302421	PRKCB
87	rs2188355	PRKCB
88	rs198145	PRKCB
89	rs195990	PRKCB
90	rs2340988	PRKCB
91	rs2636957	PRKCB
92	rs3785383	PRKCB
93	rs1976194	PRKCB
94	rs198178	PRKCB
95	rs880824	PRKCB
96	rs11074594	PRKCB
97	rs392715	PRKCB
98	rs8055243	PRKCB
99	rs7404928	PRKCB
100	rs381901	PRKCB
101	rs10852259	PRKCB
102	rs1548384	PRKCB
103	rs12445719	PRKCB
104	rs1873423	PRKCB
105	rs9928102	PRKCB
106	rs11645672	PRKCB
107	rs10492795	PRKCB
108	rs198143	PRKCB
109	rs7195728	PRKCB
110	rs1483185	PRKCD
111	rs2306572	PRKCD
112	rs3773732	PRKCD
113	rs6740453	PRKCE
114	rs4952774	PRKCE
115	rs3924523	PRKCE
116	rs281472	PRKCE
117	rs13402503	PRKCE
118	rs7604415	PRKCE

119	rs3820729	PRKCE
120	rs10191412	PRKCE
121	rs11125055	PRKCE
122	rs1947195	PRKCE
123	rs935653	PRKCE
124	rs666214	PRKCE
125	rs1375055	PRKCE
126	rs1463162	PRKCE
127	rs12185636	PRKCE
128	rs642200	PRKCE
129	rs4953274	PRKCE
130	rs542413	PRKCE
131	rs1375341	PRKCE
132	rs10495927	PRKCE
133	rs7594827	PRKCE
134	rs7582320	PRKCE
135	rs935672	PRKCE
136	rs10495928	PRKCE
137	rs1562653	PRKCE
138	rs1124787	PRKCE
139	rs4952775	PRKCE
140	rs1470598	PRKCE
141	rs6720975	PRKCE
142	rs7581914	PRKCE
143	rs629772	PRKCE
144	rs4952779	PRKCE
145	rs10175198	PRKCE
146	rs952331	PRKCE
147	rs6743144	PRKCE
148	rs4953245	PRKCE
149	rs6742737	PRKCE
150	rs2176347	PRKCE
151	rs4953262	PRKCE
152	rs4953292	PRKCE
153	rs506134	PRKCE
154	rs940052	PRKCE
155	rs483404	PRKCE
156	rs4245804	PRKCE
157	rs6719779	PRKCE
158	rs585156	PRKCE
159	rs4446102	PRKCE
160	rs951012	PRKCE
161	rs4264601	PRKCE
162	rs3768757	PRKCE
163	rs281508	PRKCE
164	rs3886870	PRKCE

165	rs3768753	PRKCE
166	rs4953251	PRKCE
167	rs2595202	PRKCE
168	rs6712557	PRKCE
169	rs6716268	PRKCE
170	rs2711292	PRKCE
171	rs7577664	PRKCE
172	rs4952787	PRKCE
173	rs2204204	PRKCE
174	rs13404973	PRKCE
175	rs14138	PRKCE
176	rs7601785	PRKCE
177	rs6751349	PRKCE
178	rs6724315	PRKCE
179	rs3745405	PRKCG
180	rs454006	PRKCG
181	rs3783778	PRKCH
182	rs14095	PRKCH
183	rs17098304	PRKCH
184	rs3742633	PRKCH
185	rs1126133	PRKCH
186	rs10141067	PRKCH
187	rs11621346	PRKCH
188	rs1033910	PRKCH
189	rs767755	PRKCH
190	rs767757	PRKCH
191	rs3783814	PRKCH
192	rs2181985	PRKCH
193	rs3783785	PRKCH
194	rs10483739	PRKCH
195	rs1091679	PRKCH
196	rs959728	PRKCH
197	rs1088672	PRKCH
198	rs1957895	PRKCH
199	rs2140825	PRKCI
200	rs1684885	PRKCI
201	rs7914917	PRKCQ
202	rs4748153	PRKCQ
203	rs6602745	PRKCQ
204	rs10508307	PRKCQ
205	rs494800	PRKCQ
206	rs596866	PRKCQ
207	rs4750439	PRKCQ
208	rs2453	PRKCQ
209	rs677986	PRKCQ
210	rs500766	PRKCQ

211	rs673964	PRKCQ
212	rs521153	PRKCQ
213	rs3793729	PRKCQ
214	rs2026431	PRKCQ
215	rs1409876	PRKCQ
216	rs3793727	PRKCQ
217	rs943452	PRKCQ
218	rs3815975	PRKCQ
219	rs6602820	PRKCQ
220	rs472317	PRKCQ
221	rs2026432	PRKCQ
222	rs658230	PRKCQ
223	rs4750531	PRKCQ
224	rs587198	PRKCQ
225	rs582052	PRKCQ
226	rs604663	PRKCQ
227	rs4357599	PRKCQ
228	rs648778	PRKCQ
229	rs501878	PRKCQ
230	rs688391	PRKCQ
231	rs11259403	PRKCQ
232	rs2236380	PRKCQ
233	rs586457	PRKCQ
234	rs7082071	PRKCQ
235	rs571715	PRKCQ
236	rs591441	PRKCQ
237	rs7072496	PRKCQ
238	rs6602747	PRKCQ
239	rs884080	PRKCZ
240	rs3107151	PRKCZ
241	rs908742	PRKCZ
242	rs3128291	PRKCZ
243	rs12755035	PRKCZ
244	rs12589992	PRKD1
245	rs7145714	PRKD1
246	rs7148794	PRKD1
247	rs875677	PRKD1
248	rs1958990	PRKD1
249	rs2273812	PRKD1
250	rs4357843	PRKD1
251	rs1953722	PRKD1
252	rs2333660	PRKD1
253	rs1958988	PRKD1
254	rs2273814	PRKD1
255	rs3783297	PRKD1
256	rs1959431	PRKD1

257	rs11626603	PRKD1
258	rs11851625	PRKD1
259	rs2300892	PRKD3
260	rs10178975	PRKD3
261	rs3770764	PRKD3
262	rs10460527	PRKD3
263	rs11896614	PRKD3
264	rs7039190	RXRA
265	rs11185659	RXRA
266	rs3132294	RXRA
267	rs4917354	RXRA
268	rs4240705	RXRA
269	rs3118571	RXRA
270	rs3818740	RXRA
271	rs10881582	RXRA
272	rs11102986	RXRA
273	rs3118536	RXRA
274	rs714112	RXRB
275	rs100537	RXRG
276	rs10800098	RXRG

Appendix 4A: Publication in Chapter 4

Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*. 2014; 85(4):399-404.

Chapter 5. Association between Multiple Sclerosis risk-associated SNPs and relapse and disability - a prospective cohort study

5.1 Preface

The previous chapter has described how to detect gene-vitamin D interactions influencing MS clinical course and particularly focus on the genes involved in the vitamin D metabolism and VDR/RXR transcript factor formation complex pathway. We have demonstrated that PKC family genes significantly modified the relationship between serum 25(OH)D and relapse in patients with MS. In this chapter, I will discuss how to detect the interactions between known MS-associated susceptibility and serum 25(OH)D influencing MS clinical course, including relapse and disability via a well-validated prospective cohort study. Most parts of this chapter have been published in *Multiple Sclerosis.* 2013 Jul 25. [Epub ahead of print] (Appendix 5A), wheras text marked as grey boxes are added and were not part of the original publication.

5.2 Introduction

To date, the human leukocyte antigen (HLA) region and more than 60 non-HLA genetic loci have been identified as being associated with MS onset. However, there has been little success in the search for genetic susceptibility to MS clinical course and disease severity, with large GWAS identifying no significant associations [1-3] apart from an earlier age of onset effect with HLA-DRB1:15*01 positivity [2,4-7]. Two large Australian cohort studies found no evidence that genetic markers including seven known MS risk-associated SNPs that were significantly associated with a number of measures of

MS progression [8,9]. Relapse-associated genetic variations have been far more difficult to study, possibly due to the narrow time-frame during which relapses occur and during which the genes may exert their effects on relapse risk, as well as the relative difficulty in assessing relapse in longitudinal studies.

There is also a possibility that many genes may realise much of their effects of disease course via gene-environment interaction. Of particular interest among the environmental factors associated with MS risk and clinical course, and therefore potentially interacting with genetic predictors of disease, are vitamin D and UV exposure. A number of studies have investigated the role of vitamin D in modulating MS clinical course, including more recent studies using longitudinal cohort designs, demonstrating an inverse association between serum 25-hydroxyvitamin D **(**25(OH)D) levels and monthly relapse rates [10], and an inverse association between 25(OH)D levels and subsequent hazard of relapse [11-13]. Gene-environment interactions between polymorphisms in the vitamin D receptor (VDR) and either childhood UVR exposure or vitamin D intake and MS onset also have been demonstrated [14,15].

Vitamin D has been shown to have significant effects on gene expression with over 2000 genes demonstrating a change in expression in the presence of increased levels of $1,25(OH)_2D$ in-vitro [16] and $1,25(OH)_2D$ has been shown to have multiple effects on the immune system largely through up-regulation of regulatory T-cells and a tendency to shift the immune system to a more tolerogenic state [17]. Recent GWAS have shown that genes that are associated with T-cell immunity are over-represented in the areas where SNPs are associated with MS onset, and have shown that these genes are significantly more likely to have vitamin D response elements and to have their expression altered by

increased 1,25$(OH)_2$D [16,18].

Given these facts, we therefore examined whether the known MS-associated genetic loci exert an effect on relapse and disability progression, and whether these loci modified the association between 25(OH)D and relapse. These associations cannot be assessed in a cross-sectional study such as a GWAS but requires longitudinal assessment of exposure and outcome to assess causality. Consequently, we have utilized the well-validated Tasmanian MS Longitudinal Study to assess the role of validated MS genetic variants on MS progression using a prospective cohort study design.

5.3 Materials and methods

5.3.1 Study design

The Southern Tasmanian Multiple Sclerosis Longitudinal Study was designed as a prospective cohort study to evaluate the role of UV exposure and 25(OH)D on the clinical course of MS [11,19]. Briefly, this study followed a cohort of 198 persons with clinically definite MS (2001 McDonald criteria [20]) living in southern Tasmania, Australia between 2002 and 2005. Of these, 145 participants with the relapsing-remitting MS (RRMS) phenotype were followed beyond one review. Of the total cohort, 188 persons had genotype data. Of the subgroup of 145 persons included in the relapse analysis, 141 had genotype data.

Where participants discontinued participation or were lost to follow-up (8/198; 4%), they were censored at the date of study exit or their last attended review. Ethics approval was obtained from the Southern Tasmania Human Research Ethics Committee. All participants provided informed consent.

5.3.2 Measurement of relapses, disability and 25(OH)D

The measurements have been described in detail elsewhere [11,21]. Briefly, relapses were defined according to established criteria [20] and were reported by phone or at biannual review. All relapse reports were validated by the study physician and study neurologist. Clinical disability, including the Expanded Disability Status Scale (EDSS), the Multiple Sclerosis Severity Score (MSSS) and the Scripps Neurologic Rating Scale (NRS), was measured each winter by a single physician.

At each biannual review, blood samples were taken. All samples were stored at -80°C and shielded from light. Free and bound serum 25(OH)D concentrations were measured using commercially available radioimmunoassays (Stillwater, Minnesota-DiaSorin Inc). Assays were performed following the conclusion of the study. Consequently, neither participants nor study personnel were aware of participants 25(OH)D concentrations during the study.

5.3.3 Genotyping, imputation and SNP selection

193 samples were genotyped including 164 MS cases as a part of the ANZgene MS GWAS using the Illumina Infinium Hap370CNV Array [22], and 29 cases were genotyped using the Illumina HumanOmniExpress-12v1_A array. All the samples were previously identified as European descent [22]. 61 known loci with genome-wide significant association with MS [1,3,14,15,22,23] were selected. Since there was a small degree of disparity in the SNPs typed between the two genotyping arrays, missing genotypes were imputed to a common panel of SNPs using the BEAGLE package [24] with 1000 Genomes data of European ancestry [25] as the reference, after a conservative quality control conducted with PLINK [26]: individuals with call rates less than 0.90, SNPs with call rates less than 0.95 or in Hardy-Weinberg equilibrium ($p<10^{-7}$), or

duplicates were excluded, leaving 189 cases with 290,536 SNPs.

Methods note 5.1 Missing genotyped SNPs imputation

As mentioned above, there was a small degree of disparity in the SNPs typed between the two genotyping arrays, missing genotypes need to impute into a common panel of SNPs using the BEAGLE package [24] with 1000 Genomes data of European ancestry [25] as the reference. To well understand the imputation, more details were added here. Firstly, the merged and cleaned genotypes dataset was converted to BEAGLE format by using the linkage2Beagle.jar utility program. We then changed the genotype data as the same human genome version (e.g. hg 19), since 1000 Genomes data refer to hg19, and the position, alleles and order for each SNP may be different from two human genome versions (hg 18 vs. hg19). By editing Perl program, we changed the SNP position and flipped the alleles, and then we used Beagle "check-strand" utility program to get the chromosomal locations refer to hg19 version. Finally we used the BEAGLE method for phasing and imputing the data simultaneously by using the high performance computing facilities in University of Tasmania (The Tasmanian Partnership for Advanced Computing (TPAC). The output includes the allele dosage of each SNP for each individual, the imputation score (R^2) for each marker and the phased files. Those genotyped or imputed SNPs can be assessed with imputation score (R^2). For example, the genotyped SNPs' imputation scores should equal to 1, and among the imputed SNPs, those with imputation score closer to 1, the higher likelihood that it imputed from the nearby genotyped SNPs, and then more reliable to represent the missing genotyped SNP. When one SNP's imputation score less than 0.5, generally it should be excluded for the analysis.

Finally, one non-HLA locus (rs17445836 near *IRFB* gene) was excluded due to a low imputation score (R^2<0.5), leaving 60 known non-HLA MS-associated loci for the analysis. Additionally single SNP genotyping for HLA-DRB1*15:01 has been previously reported for this cohort with 133 of 141 cases having this genotyping data.

5.3.4 Measurement of covariates

Height and weight were measured at study entry, and body mass index (BMI) was calculated as weight divided by height squared. At entry and each biannual review, participants were asked about their lifestyle, including time in the sun during weekends and holidays in the current season and previous season, physical activity, smoking, fish intake, immunomodulatory therapy use, and vitamin D supplement use. Skin melanin density was measured on the upper inner arm using a spectrophotometer [27].

5.3.5 Data analysis

The hazard of relapse was modeled as a function of 25(OH)D and other covariates on time-to-relapse and was evaluated using Cox proportional hazards models for repeated events, using the gap-time model described by Prentice and colleagues [28]. Standard errors were adjusted to reflect multiple events per person. All covariates satisfied the proportional hazards assumption with the exception of the binary variable for sex and the categorical variable for baseline EDSS (0 – 2.5, 3.0 – 5.0, 5.5– 7.0, 7.5 – 9.0). For this reason, all models are stratified to allow the baseline hazards to differ by sex and baseline EDSS category (0 – 2.5, 3.0 – 5.0, 5.5– 7.0, 7.5 – 9.0). 25(OH)D was estimated at

monthly intervals between biannual measurements using methods described previously [11]. Briefly, a sinusoidal curve was fitted to the measured 25(OH)D and used to predict values at monthly intervals between measures. To examine whether there was an interaction between 25(OH)D and a SNP on relapse, a product term was included in the model. To examine whether there was an interaction between modeled 25(OH)D and a SNP on relapse, a product term was included in the model, and then the association between 25(OH)D and hazard of relapse can be estimated at each SNP allele level.

Predictors of 25(OH)D were evaluated by multilevel mixed-effects linear regression, with SNPs as the predictors and 25(OH)D as the outcome adjusted for season of measure (summer, winter), age, BMI, melanin density (%), self-reported sun exposure ("less than 0.5 hour a day; "1/2 to 1 hour a day"; "1 to 2 hours a day"; "2 to 3 hours a day"; "3 to 4 hours a day"; greater than 4 hours a day"), vitamin D supplementation ("never" ; "less than 200 IU a day"; "200 to 400 IU a day"), fish intake ("never"; "less than once a week"; "1 to 2 serves a week"; "2 to 4 serves a week"; "greater than 4 serves a week") and smoking (No, Yes). Dependent variables were transformed as required to make the residuals less heteroskedastic; however all regression coefficients are reported on the scale of the original variable.

To estimate cumulative effects of the significant SNPs on relapse or on levels of 25(OH)D, we created a variable that provided values for the number of risk genotypes affecting relapse or for the number of genotypes associated with low 25(OH)D to represent the "genetic risk score". For example, those subjects with 3, 4 or 5 genotypes that associated with lower 25(OH)D levels were each compared with the reference group - those carrying less than 3 genotypes associated with lower 25(OH)D levels.

All data analyses were performed using STATA/IC V.12.1 (StataCorp LP, College Station, Texas, USA).

5.4 Results

The cohort of 188 persons with genotype data was followed for an average of 2.3 years. The cohort was 68.6% female and of mean age 48.5 years (SD: 11.4). Of those with RRMS at study entry who were followed beyond one review and had genotype data (n=141), 75.2% were female and of mean age 45.8 years (SD: 10.4). A total of 122 confirmed relapses occurred in 70 participants.

We identified five SNPs (rs11810217, rs2760524, rs1790100, rs2300603, rs2283792) in the genes *EVI5, RGS1, MPHOSPH9, BATF, MAPK1,* respectively, which were associated with the hazard of relapse (**Table 5.1**), these associations persisted after mutual adjustment and after adjustment for immunomodulatory medication use (data not shown). In addition, we found five SNPs (rs11581062, rs7595037, rs17824933, rs180515, rs874628) in five genes (*VCAM1, PLEK, CD6, RPS6KB1*, *MPV17L2,* respectively) that were associated with the levels of 25(OH)D, even after adjustment for sun exposure and other confounders. We also identified three SNPs *(*rs354033, rs2119704, rs2248359)in the genes *ZNF767, GALC, CYP24A1* that modified the relationship between 25(OH)D and the hazard of relapse, even after mutual adjustment and after adjustment for immunomodulatory medication use (data not shown). None of these associations remained significant after adjustment for multiple comparisons for the 60 loci, however (**Table 5.1**). HLA-DRB1*15:01 was not associated with hazard of relapse (p_{trend} =0.64)

nor 25(OH)D levels (p_{trend} =0.55), and had no modulating effect on the relationship between 25(OH)D and relapse ($p_{interaction}$ =0.33).

Table 5.1. Significant SNPs and their associations with hazard of relapse, levels of 25(OH)D and the 25(OH)D-relapse association

5 SNPs associated with the hazard of relapse

Chr	SNP	bp	Allele	MAF	Genotyped / imputed	HR_relap (95%CI)	*P_relap*	*Padj_relap*	*P_interact*	*P_25(OH)D*	Nearest Gene	Gene function*
1	rs11810217[20]	92920965	C/T	0.385	imputed	1.46(1.10, 1.93)	0.008	0.48	0.92	0.95	EVI5	A regulator of cell cycle progression
1	rs2760524[21]	190797171	A/G	0.126	imputed	0.60(0.42, 0.86)	0.005	0.30	0.14	0.64	RGS1	Inhibits signal transduction; regulates B-cell activation and proliferation
12	rs1790100[21]	122222678	T/G	0.218	imputed	1.38(1.02, 2.09)	0.039	1.00	0.68	0.80	MPHOSPH9	Unknown
14	rs2300603[20]	75075310	C/T	0.233	genotyped	1.45(1.00, 2.09)	0.049	1.00	0.64	0.72	BATF	A negative regulator of AP-1
22	rs2283792[20]	20461125	T/G	0.477	imputed	1.44(1.12, 1.87)	0.005	0.30	0.08	0.79	MAPK1	A transcriptional repressor; involved in cellular processes.

5 SNPs associated with the levels of 25(OH)D (nmol/L)

Chr	SNP	bp	Allele	MAF	Genotyped / imputed	Coef_25(OH)D (95%CI)	*P_25(OH)D*	*Padj_25(OH)D*	*P_interact*	*P_relap*	Nearest Gene	Gene function*
1	rs11581062[20]	101180107	G/A	0.341	genotyped	+4.33(0.44, 8.21)	0.029	1.00	0.35	0.30	VCAM1	Leukocyte-endothelial cell adhesion
2	rs7595037[20]	68500599	C/T	0.439	genotyped	-3.16(-5.99, -0.32)	0.029	1.00	0.43	0.87	PLEK	Major protein kinase C substrate of platelets
17	rs180515[20]	55379057	G/A	0.392	imputed	+3.79(0.09, 7.50)	0.045	1.00	0.42	0.17	RPS6KB1	Integrate nutrient and growth factor signals
11	rs17824933[21]	60517188	G/C	0.272	imputed	+5.68(1.40, 9.97)	0.009	0.54	0.18	0.74	CD6	Involved in cell adhesion
19	rs874628[20]	18165700	G/A	0.273	genotyped	-3.99(-7.27, -0.70)	0.028	1.00	0.35	0.14	MPV17L2	Unknown

3 SNPs modified the relationship between 25(OH)D and the hazard of relapse

Chr	SNP	bp	Allele	MAF	Genotyped / imputed	HR_25(OH)D (95%CI)	*P_interact*	*Padj_interact*	*P_relap*	*P_25(OH)D*	Nearest Gene	Gene function*
7	rs354033[20]	148920397	A/G	0.262	imputed	0.88(0.78, 0.99)	0.029	1.00	0.28	0.41	ZNF767	Unknown
14	rs2119704[20]	87557442	A/C	0.103	imputed	1.33(1.07, 1.65)	0.011	0.66	0.94	0.95	GALC	Hydrolyzes galactose ester bonds
20	rs2248359[20]	52224925	T/C	0.384	genotyped	1.14(1.03, 1.27)	0.015	0.90	0.35	0.84	CYP24A1	Maintain calcium homeostasis. Catalyzes 25(OH)D and $1,25(OH)_2D$

P_relap: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons for 60 loci; *Padj_relap*: further adjusted for multiple comparisons for 60 loci.

P_25(OH)D: adjusted for sun exposure, season, age, BMI, vitamin D supplementation, melanin density, fish intake and smoking, before adjustment for multiple comparisons for 60 loci. *Padj_25(OH)D*: further adjusted for multiple comparisons for 60 loci.
P_interact: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons for 60 loci; *Padj_interact*: further adjusted for multiple comparisons for 60 loci.

All the SNPs with imputation R^2 >0.9; rs2300603, rs11581062, rs7595037, rs874628, rs2248359 are genotype SNPs (R^2 =1).

*Gene function index from UCSC Browser: http://genome.ucsc.edu/

[20]= original from WTCCC2 MS GWAS; [21]= original from meta-analysis MS GWAS.

Bp =base pair (hg18); HR=hazard ratio; 25(OH)D= 25-hydroxyvitamin D; $1,25(OH)_2D$= 1,25-dihydroxyvitamin D; CI= confidence interval.

Despite the absence of statistical significance after adjustment for multiple comparisons, a clear allele dose-response was observed for two SNPs (rs11810217 tagging *EVI5*, and rs2283792 tagging *MAPK1*) that were associated with the hazard of relapse (**Table 5.2).** When we examined the combined effect on relapse of these five relapse-associated SNPs, we found a clear dose-response effect with increasing number of risk genotypes (p_{trend}=0.00013). For example, compared to subjects carrying less than 2 risk genotypes, those with 4 or 5 risk genotypes had hazard ratios of 2.22 (95% CI: 1.36-3.60) and 2.37 (95% CI: 1.35-4.15), respectively (**Table 5.2**).

Table 5.2. The allele dosage response and cumulative effect for the 5 SNPs that were associated with the hazard of relapse

Chr/gene	SNP	Genotype	N	HR_relap (95%CI)	*p*
1/ EVI5	rs11810217	TT	19	1.00 [Reference]	
		TC	50	1.74(0.92, 3.29)	0.088
		CC	72	2.44(1.33,4.47)	0.004
					p_{trend} =0.008
1/ RGS1	rs2760524	GG	100	1.00 [Reference]	
		AA+AG	41	1.82(1.21, 2.72)	0.004
12/ MPHOSPH9	rs1790100	TT	80	1.00 [Reference]	
		GG+GT	61	1.50(1.01, 2.24)	0.044
14/ BATF	rs2300603	CC+CT	57	1.00 [Reference]	
		TT	84	1.54(1.02, 2.34)	0.041
22/ MAPK1	rs2283792	GG	41	1.00 [Reference]	
		GT	69	1.93(1.09, 3.43)	0.024
		TT	31	2.16(1.18, 3.95)	0.013
					p_{trend} =0.005
Cumulative risk of relapse for the 5 SNPs					
<=2 risk genotypes*			73	1.00 [Reference]	
3 risk genotypes			26	1.24(0.68, 2.26)	0.49
4 risk genotypes			27	2.22(1.36, 3.60)	0.001
5 risk genotypes			15	2.37(1.35, 4.15)	0.003
					p_{trend} =0.00013

* None carried 0 risk genotype, 1 person carried 1 risk genotype.

p trend: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

The homozygous genotypes were combined with the heterozygous ones because of small numbers (AA genotype of rs2760524 (n=2); GG genotype of rs1790100 (n=7); CC genotype of rs2300603 (n=4)), and were excluded for *p* trend test.

A clear allele dose-response was also observed for three of the five SNPs that associated with 25(OH)D serum levels (rs11581062, rs7595037, rs180515), such that with the subjects homozygous for the minor allele had significantly higher 25(OH)D levels relative to those homozygous for the common allele. When we examined the combined effect on 25(OH)D levels of those five 25(OH)D-associated SNPs ((rs11581062, rs7595037, rs180515, rs17824933, rs874628), a clear dose-response was observed with increasing number of genotypes that were associated with lower 25(OH)D (p_{trend}= 1.03x10^{-5}). For example, compared to those carrying 2 'risk' genotypes, those with 4 or 5 genotypes had 25(OH)D levels that were 7.96 nmol/L ((95% CI: (-14.89, -1.02); p=0.024)) and 13.54 nmol/L lower ((95% CI: (-20.52, -6.56); p=0.00014)), respectively (**Table 5.3**).

Table 5.3. The allele dosage response and the cumulative effect of the 5 SNPs that were associated with 25(OH)D

Chr/gene	SNP	Genotype	N	25(OH)D (95%CI)	*p*
1/ VCAM1	rs11581062	AA	62	48.55(45.13,51.97)	ref
		AG	63	+2.40(-2.56, 7.35)	0.343
		GG	16	+10.46(1.63,19.29)	0.02
					p_{trend} =0.029
2/ PLEK	rs7595037	TT	53	48.33(44.66, 52.01)	ref
		TC	52	+1.86(-3.48, 7.19)	0.495
		CC	36	+6.87(0.56, 13.18)	0.033
					p_{trend} =0.029
7/ RPS6KB1	rs180515	AA	55	48.64(45.01, 52.26)	ref
		AG	65	+2.10(-2.92, 7.12)	0.412
		GG	21	+8.56(0.73, 16.40)	0.032
					p_{trend} =0.045
11/ CD6	rs17824933	CC	76	47.66(44.71, 50.62)	ref
		CG+GG	65	+6.99(2.25, 11.73)	0.004
					p trend =0.009
19/ MPV17L2	rs874628	GG	14	46.07(39.22, 52.91)	ref
		GA	52	+2.22(-5.67, 10.10)	0.58
		AA	75	+7.22(-0.41, 14.86)	0.06
					p_{trend} =0.028

Cumulative effects of the 5 SNPs associated with lower 25(OH)D			
<=2 genotypes*	28	57.53(51.79, 63.27)	ref
3 genotypes	31	-2.55(-10.27, 5.17)	0.517
4 genotypes	47	-7.96(-14.89, -1.02)	0.024
5 genotypes	35	-13.54(-20.52, -6.56)	0.00014
			p_{trend} =1.03 x10^{-5}

*None carried 0 and 1 genotype.

p trend: adjusted for sun exposure, season, age, BMI, vitamin D supplementation, melanin density, fish intake and smoking, before adjustment for multiple comparisons.
The homozygous genotypes were combined with the heterozygous ones because of small numbers (GG genotype of rs17824933 (n=8))

Among the three SNPs that modified the relationship between 25(OH) and relapse, for rs354033 tagging *ZNF767* and rs2248359 tagging *CYP24A1*, individuals carrying two copies of the common allele showed an inverse relationship between 25(OH)D and relapse, while those carrying at least one copy of the rare allele showed no significant 25(OH)D-relapse association (**Table 5.4**).

Table 5.4. The association between 25(OH)D and hazard of relapse stratified by allele for the 3 SNPs that modified the 25(OH)D-relapse association

Chr/gene	SNP	Genotype	N	HR_25(OH)D (95%CI)	*p*
7/ ZNF767	rs354033	GG	79	0.81(0.73, 0.88)	7.46x10^{-6}
		AA+AG	62	0.95(0.85, 1.06)	0.372
					$p_{interaction}$ =0.029
14/ GALC	rs2119704	CC	116	0.90(0.83, 0.97)	0.006
		AA+AC	25	0.68(0.55, 0.82)	0.0002
					$p_{interaction}$ =0.011
20/ CYP24A1	rs2248359	CC	50	0.78(0.70, 0.86)	3.48x10^{-6}
		CT	71	0.91(0.80, 1.03)	0.146
		TT	20	0.99(0.81, 1.17)	0.922
					$p_{interaction}$ =0.015

p interaction: adjusted for age, sex and baseline EDSS, before adjustment for multiple comparisons.

The homozygous genotypes were combined with the heterozygous ones because of small numbers (AA genotype of rs354033 (n=7); AA genotype of rs2119704 (n=1)).

We also evaluated the association between the 60 loci and cross-sectional measured disability or progression in disability, but found no consistent evidence for an association between any SNPs and progression in disability (**Table 5.5)**

Table 5.5. Top SNPs and their associations with progression in clinical disability

Chr/gene	SNP	Genotype	N	Baseline MSSS		Annual change in EDSS		Annual change in NRS	
				MSSS (95% CI)	*p*	Coef. (95% CI)	*p*	Coef (95% CI)	p
1/CD58	rs1335532	AA	160	4.57(4.17,4.97)	ref	0.00(ref)	ref	0.00(ref)	ref
		AG	25	+0.53(-0.57,1.62)	0.35	+0.12(-0.07,0.31)	0.22	-0.28(-2.31,1.74)	0.79
		GG	4	+1.21(-1.49,3.90)	0.38	+0.34(-0.03,0.71)	0.07	+0.39(-2.25,3.04)	0.77
	Trend				*p*=0.21		*p*=0.044		*p*=0.91
5/PTGER4	rs4613763	CC	5	4.91(2.65,7.16)	ref	0.00 (ref)	ref	0.00(ref)	ref
		CT	41	-0.69(-3.05,1.67)	0.56	+0.14(-0.07,0.35)	0.19	+1.25(-1.59,4.08)	0.39
		TT	140	-0.08(-2.38,2.22)	0.95	+0.10(-0.07,0.26)	0.25	-0.84(-3.50,1.83)	0.54
	Trend				*p*=0.30		*p*=0.87		*p*=0.008
5/IL12B	rs2546890	AA	54	4.74(4.05,5.43)	ref	0.00 (ref)	ref	0.00(ref)	ref
		AG	92	+0.14(-0.73,1.02)	0.75	+0.06(-0.10,0.22)	0.49	+0.50(-0.89,1.88)	0.48
		GG	43	-0.59(-1.61,0.43)	0.26	-0.08(-0.26,0.10)	0.36	+1.67(0.04,3.30)	0.045
	Trend				*p*=0.30		*p*=0.44		*p*=0.049
6/THEMIS	rs802734	AA	91	4.43(3.90,4.95)	ref	0.00 (ref)	ref	0.00(ref)	ref
		AG	80	+0.40(-0.37,1.18)	0.31	+0.05(-0.08,0.18)	0.46	-1.36(-2.60,-0.12)	0.032
		GG	18	+0.78(-0.58,2.13)	0.26	+0.28(0.01,0.55)	0.043	-1.44(-3.38,0.51)	0.147
	Trend				*p*=0.18		*p*=0.06		*p*=0.029
8/MYC	rs4410871	CC	108	4.43(3.96,4.90)	ref	0.00 (ref)	ref	0.00(ref)	ref
		CT	70	+0.64(-0.15,1.43)	0.11	-0.01(-0.15,0.13)	0.85	+1.28(0.09,2.48)	0.035
		TT	11	+0.10(-1.52,1.72)	0.91	-0.17(-0.42,0.07)	0.17	+1.22(-1.24,3.68)	0.33
	Trend				*p*=0.26		*p*=0.36		*p*=0.048
19/MPV17L2	rs874628	AA	103	4.63(4.13,5.12)	ref	0.00 (ref)	ref	0.00(ref)	ref
		AG	69	-0.002(-0.81,0.80)	1.00	+0.06(-0.08,0.21)	0.37	+0.42(-0.93,1.77)	0.54
		GG	17	+0.46(-0.88,1.80)	0.50	+0.29(0.02,0.57)	0.033	-2.52(-4.61,-0.42)	0.019
	Trend				*p*=0.62		*p*=0.032		*p*=0.18
20/CYP24A1	rs2248359	CC	71	4.88(4.27,5.50)	ref	0.00 (ref)	ref	0.00(ref)	ref
		CT	91	-0.42(-1.23,0.39)	0.31	+0.04(-0.10,0.18)	0.56	-0.54(-1.78,0.71)	0.40
		TT	27	-0.07(-1.23,1.09)	0.91	+0.27(0.08,0.45)	0.004	-0.05(-2.08,1.99)	0.96
	Trend				*p*=0.64		*p*=0.018		*p*=0.74
22/MAPK1	rs2283792	GG	41	4.31(3.60,5.01)	ref	0.00 (ref)	ref	0.00(ref)	ref
		GT	69	+0.51(-0.36,1.39)	0.25	+0.07(-0.07,0.21)	0.35	-0.61(-1.91,0.69)	0.36
		TT	31	+0.47(-0.58,1.52)	0.38	+0.14(-0.07,0.36)	0.19	-1.80(-3.56, -0.04)	0.045
	Trend				*p*=0.37		*p*=0.18		*p*=0.046

5.5 Discussion

Using a well-validated prospective MS cohort designed to assess the effects of vitamin D and/or personal UVR exposure on MS clinical course, we investigated whether known MS risk-associated SNPs were associated with relapse, disability progression and 25(OH)D levels, and whether these SNPs modified the relationship between 25(OH)D and relapse. We found five SNPs in or near genes *EVI5, RGS1, MPHOSPH9, BATF, MAPK1* that were associated with relapse, where the risk alleles associated with MS onset also predicted increased relapse risk [23]. We also found three SNPs in or near genes *ZNF767, GALC, CYP24A1* that modified the relationship between 25(OH)D and the hazard of relapse, and five SNPs in or near genes *VCAM1, PLEK, CD6, RPS6KB1, MPV17L2* that were found associated with the levels of 25(OH)D. However, no SNP association remained significant after adjustment for multiple comparisons.

It is not surprising that our primary results were not significant after adjusting for multiple comparisons, given the number of participants required and/or the length of follow-up required to reach that level of significance is prohibitively difficult for studies of MS clinical course. Therefore we sought further evidence of association utilising allele dose-responses and cumulative genotype effects to provide support for our findings. These findings support the nominal SNP associations and provide evidence for associations, despite their not meeting the multiple testing correction thresholds, and thus may support an interpretation of a likely association. For example, two relapse-associated SNPs (rs11810217, rs2283792) and three 25(OH)D-associated SNPs (rs11581062, rs7595037, rs180515) showed a clear allele-response. Furthermore, when studying the

effects of risk SNPs on 25(OH)D levels, compared to those carrying 2 'risk' genotypes, those with 4 or 5 'risk' genotypes had 25(OH)D levels that were 7.96 nmol/L ((95% CI: (-14.89, -1.02); p=0.024)) and 13.54 nmol/L lower ((95% CI: (-20.52, -6.56); p=0.00014)), respectively. Similarly, regarding the hazard of relapse, compared to subjects carrying less than 2 risk genotypes, those with 4 or 5 risk genotypes had hazard ratios of 2.22 (95% CI: 1.36-3.60) and 2.37 (95% CI: 1.35-4.15), respectively. Both of these findings are of clinical significance. Despite this, those SNPs associated with 25(OH)D levels were not associated with relapse or modified the 25(OH)D-relapse association, possibly because the magnitude of effect on 25(OH)D levels was too small in order to mediate an effect on relapse.

Our findings do not support an association between the genetic loci studied and any of the measures of disability progression. We hypothesised that disability progression and relapse have differing genetic susceptibilities, with MS relapse driven by acute inflammation and disability more by neurodegeneration. As nearly all of the identified MS risk SNPs tagging genes involved in immunological/inflammatory pathways, their association with relapse but not disability progression is consistent with this construction.

One theory of MS causation is that it is the result of a misdirected immune response by autoreactive T-cells against as yet undetermined CNS antigens [29,30] supported by the finding that MS genetic determinants are heavily biased towards variants in T-cell genes and the HLA region. One variant we found to be associated with relapse in our cohort was located in the *RGS1* genes, which have been demonstrated to have important roles in the immune system [3], with *RGS1* involved in T-cell signal transduction [31]. Furthermore, the associations between lower levels of 25(OH)D and relapse are likely mediated by the

immunomodulatory effects of microenvironmental $1,25(OH)_2D$, acting to up-regulate regulatory T-cell function and depress inflammatory immune activity [17]. In line with previous evidence that did not show HLA DRB1*15:01 to be associated with MS clinical course [3], in our cohort HLA DRB1*15:01 status was not associated with relapse nor did it modulate the association between 25(OH)D levels and relapse. This despite the fact that HLA DRB1*15:01 is known to be upregulated by $1,25(OH)_2D$ through the well-described VDRE in its promoter region [32]. Interestingly, among those genes associated with 25(OH)D levels or which modulated the relationship between relapse and 25(OH)D in our cohort, *ZNF767* and *VCAM* bind VDR and *CYP24A1* also demonstrates VDR binding and is a core component of 25(OH)D metabolism [18,33,34]. Thus these findings provide some support for our initial hypothesis that the effect of vitamin D on relapse may be partly modulated by genetic variants in established MS-associated genes.

A core strength of our study is its prospective methodology and the availability of detailed environmental data, including potential confounders such as sun exposure and vitamin D supplementation. A weakness of our study is the sample size. Although this is one of the largest and most well-studied MS cohorts for the study of MS clinical course available, the limitation imposed by the sample size reflects the difficulty of undertaking studies of this nature. Genetic associations with clinical course can best be answered in longitudinal studies such as ours, but multiple testing penalties associated with studying large groups of genes and/or tagging SNPs requires very large sample size and/or follow-up. Therefore, we have used other methodologies to overcome potential type-1 error, including allele dose responses and accumulated genotype risk scores, as well as consistency between outcome measures (multiple measures of disability). Therefore we believe that in studies of this type where sample size effects significantly hinders the

ability to detect statistically significant genetic associations, the utilisation of this ancillary supporting evidence to provide additional support is a worthwhile analysis method. These associations can then be studied in other cohorts as *a priori* hypotheses without the need for multiple testing limitations. Our results, in combination with other studies, can bridge the gap between the need for larger and larger genetic studies and the realities of undertaking MS longitudinal studies with careful real-time verification of relapse history.

In conclusion, our prospective cohort study provides evidence that MS risk-associated SNPs exert an effect on MS relapse but not progression in disability. In addition, some SNPs were associated with serum 25(OH)D levels and modified the relationship between vitamin D and relapse. These findings support the hypothesis that gene-environment interactions may be an important mechanism by which MS clinical course is driven, and provide further support for the role of vitamin D in MS relapse.

5.6 Summary

Background: The modulating effects of the MS risk-associated SNPs on MS clinical course are not well established.

Objectives: To investigate whether known MS risk-associated SNPs were associated with clinical course, and whether these SNPs modified the 25(OH)D-relapse association.

Methods: Using a prospective cohort of 141 participants with relapsing-remitting MS and genotype data followed between 2002 and 2005, genotype-vitamin D interactions and the genetic predictors of relapse were assessed using survival analysis and genetic predictors of 25(OH)D and disability progression were evaluated by multilevel mixed-effects linear regression.

Results: While no SNP reached statistical significance after multiple testing, five SNPs were associated with relapse, with significant cumulative genotype risk effects and two demonstrated significant allele dose-response. Two SNPs altered the 25(OH)D-relapse association with significant allele dose-response. Five SNPs modified levels of 25(OH)D, with significant cumulative genotype 'risk' effect, and three demonstrated significant allele dose-response. We found no consistent evidence for an association between any SNPs and disability.

Conclusions: Our study provides evidence for an association between known MS risk-associated SNPs and relapse. Our findings indicate gene-environment interactions may be an important mechanism on MS clinical course, and provide support for the role of vitamin D in MS relapse.

5.7 Postscript

This chapter has discussed how to detect the interactions between known MS-associated susceptibility and vitamin D influencing clinical course of MS. We provided evidence that known MS-associated susceptibility also modified the relationship between vitamin D and relapse in patients with MS. In addition, these findings provide further support for the role of vitamin D in MS relapse. Next chapter we will specifically focus on interferon-beta (IFN- β), the best therapeutic agent so far which can reduce relapse rate and accumulation disability of MS, to examine whether there are genetic effects modulate IFN- β effect on serum vitamin D levels by using the same prospective cohort as the previous two chapters described.

5.8 References

1. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Human molecular genetics 18: 767-778.
2. Martinelli-Boneschi F, Esposito F, Brambilla P, Lindstrom E, Lavorgna G, et al. (2012) A genome-wide association study in progressive multiple sclerosis. Mult Scler.
3. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
4. McDonnell GV, Mawhinney H, Graham CA, Hawkins SA, Middleton D (1999) A study of the HLA-DR region in clinical subgroups of multiple sclerosis and its influence on prognosis. J Neurol Sci 165: 77-83.
5. Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, et al. (2006) Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 15: 2813-2824.
6. Okuda DT, Srinivasan R, Oksenberg JR, Goodin DS, Baranzini SE, et al. (2009) Genotype-Phenotype correlations in multiple sclerosis: HLA genes influence disease severity inferred by 1HMR spectroscopy and MRI measures. Brain 132: 250-259.
7. Cree BA, Reich DE, Khan O, De Jager PL, Nakashima I, et al. (2009) Modification of Multiple Sclerosis Phenotypes by African Ancestry at HLA. Arch Neurol 66: 226-233.
8. van der Walt A, Stankovich J, Bahlo M, Taylor BV, van der Mei IA, et al. (2009) Apolipoprotein genotype does not influence MS severity, cognition, or brain atrophy. Neurology 73: 1018-1025.
9. Jensen CJ, Stankovich J, Van der Walt A, Bahlo M, Taylor BV, et al. (2010) Multiple sclerosis susceptibility-associated SNPs do not influence disease severity measures in a cohort of Australian MS patients. PLoS One 5: e10003.
10. Tremlett H, van der Mei IA, Pittas F, Blizzard L, Paley G, et al. (2008) Monthly ambient sunlight, infections and relapse rates in multiple sclerosis. Neuroepidemiology 31: 271-279.
11. Simpson S, Jr., Taylor B, Blizzard L, Ponsonby AL, Pittas F, et al. (2010) Higher 25-hydroxyvitamin D is associated with lower relapse risk in multiple sclerosis. Ann Neurol 68: 193-203.
12. Mowry EM, Krupp LB, Milazzo M, Chabas D, Strober JB, et al. (2010) Vitamin D status is associated with relapse rate in pediatric-onset multiple sclerosis. Ann Neurol 67: 618-624.
13. Runia TF, Hop WC, de Rijke YB, Buljevac D, Hintzen RQ (2012) Lower serum vitamin D levels are associated with a higher relapse risk in multiple sclerosis. Neurology 79: 261-266.
14. Simon KC, Munger KL, Xing Y, Ascherio A (2010) Polymorphisms in vitamin D metabolism related genes and risk of multiple sclerosis. Mult Scler 16: 133-138.
15. Dickinson JL, Perera DI, van der Mei AF, Ponsonby AL, Polanowski AM, et al. (2009) Past environmental sun exposure and risk of multiple sclerosis: a role for the Cdx-2 Vitamin D receptor variant in this interaction. Mult Scler 15: 563-570.
16. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, et al. (2010) A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res 20: 1352-1360.

17. Ascherio A, Munger KL, Simon KC (2010) Vitamin D and multiple sclerosis. Lancet Neurol 9: 599-612.
18. Disanto G, Sandve GK, Berlanga-Taylor AJ, Ragnedda G, Morahan JM, et al. (2012) Vitamin D receptor binding, chromatin states and association with multiple sclerosis. Hum Mol Genet.
19. Pittas F, Ponsonby AL, van der Mei IA, Taylor BV, Blizzard L, et al. (2009) Smoking is associated with progressive disease course and increased progression in clinical disability in a prospective cohort of people with multiple sclerosis. J Neurol 256: 577-585.
20. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, et al. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50: 121-127.
21. Simpson S, Jr., Taylor B, Dwyer DE, Taylor J, Blizzard L, et al. (2012) Anti-HHV-6 IgG titer significantly predicts subsequent relapse risk in multiple sclerosis. Mult Scler 18: 799-806.
22. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
23. Patsopoulos NA, de Bakker PI, Esposito F, Reischl J, Lehr S, et al. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol 70: 897-912.
24. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210-223.
25. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, et al. (2012) The 1000 Genomes Project: data management and community access. Nature Methods 9: 459-462.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.
27. van der Mei IA, Ponsonby AL, Dwyer T, Blizzard L, Simmons R, et al. (2003) Past exposure to sun, skin phenotype, and risk of multiple sclerosis: case-control study. Bmj 327: 316.
28. Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68: 373-379.
29. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG (2000) Multiple sclerosis. N Engl J Med 343: 938–952.
30. Sosa RA, Forsthuber TG (2011) The critical role of antigen-presentation-induced cytokine crosstalk in the central nervous system in multiple sclerosis and experimental autoimmune encephalomyelitis. J Interferon Cytokine Res 31: 753-768.
31. Lin R, Charlesworth J, van der Mei I, Taylor BV (2012) The genetics of multiple sclerosis. Practical Neurology 12: 279-288.
32. Ramagopalan SV, Maugeri NJ, Handunnetthi L, Lincoln MR, Orton SM, et al. (2009) Expression of the multiple sclerosis-associated MHC class II Allele HLA-DRB1*1501 is regulated by vitamin D. PLoS Genet 5: e1000369.
33. Zierold C, Darwish HM, DeLuca HF (1995) Two vitamin D response elements function in the rat 1,25-dihydroxyvitamin D 24-hydroxylase promoter. J Biol Chem 270: 1675-1678.

34. Ohyama Y, Ozono K, Uchida M, Yoshimura M, Shinki T, et al. (1996) Functional assessment of two vitamin D-responsive elements in the rat 25-hydroxyvitamin D3 24-hydroxylase gene. J Biol Chem 271: 30381-30385.

Appendix 5A: Publication in Chapter 5

Rui Lin, Bruce V Taylor, Steve Simpson, Jr., Jac Charlesworth, Anne-Louise Ponsonby, Fotini Pittas, Terence Dwyer, Ingrid A F van der Mei. Association between multiple sclerosis risk-associated SNPs and relapse and disability - a prospective cohort study. *Multiple Sclerosis.* 2014; 20(3):313-21.

Chapter 6. Modulating effects of the gene *WT1* on the interferon-β-vitamin D association in MS

6.1 Preface

The previous two chapters demonstrated that gene-vitamin D interactions influenced clinical course, and provided further support for the role of vitamin D in MS relapse. Interferon-β (IFN-β) was the first therapeutic agent found that can significantly reduce relapse rate and probably prevent accumulation of disability in patients with MS. We have demonstrated previously that IFN-β treatment was associated with higher serum vitamin D levels among persons with MS. Therefore, we were interested in whether there are genetic effects modulating IFN-β effect on serum vitamin D levels. In this chapter, I will particularly discuss whether those genes involved in the vitamin D pathway modulate the demonstrated relationship between serum vitamin D and IFN-β, the relationship between IFN-β and sun in predicting serum vitamin D, and the interaction between IFN-β and serum vitamin D in modulating relapse risk in people with MS. This chapter is submitted.

6.2 Introduction

Interferon-β is an immunomodulatory agent that has been widely used in the treatment of relapsing-remitting MS (RRMS). Evidence has shown IFN-β treatment significantly reduces relapse rate and the occurrence of MRI-detected lesions [1]. Despite its ubiquity in treating MS, the therapeutic mechanism of IFN-β is unclear, but is thought to act via some immune-modulating mechanism, including inhibiting T-cell proliferation [2],

reduced antigen presentation [3], reduced T-cell migration into the CNS and altered cytokine production [4]. Recently, we demonstrated that IFN-β treatment was associated with higher 25-hydroxyvitamin D (25(OH)D) among persons with MS, this apparently mediated by a stronger association between personal sun exposure and subsequently measured 25(OH)D [5]. Additionally, a positive interaction was found between IFN-β treatment and sufficient 25(OH)D, such that each only exerted a therapeutic effect on relapse risk in the presence of the other, which indicated the therapeutic effects of IFN-β may involve vitamin D metabolism. Our goal here was to assess whether any genetic variants significantly modified this association, with a goal to identifying potential biological pathways underlying the effect of IFN-β on vitamin D and on relapse.

Of the environmental factors identified as modulating MS risk and clinical course, vitamin D is among the strongest and most consistent [6-10]. In addition to the effects of environment and behavior, circulating vitamin D levels have been found to be influenced by genetic determinants, as some genes involved in vitamin D metabolism have been identified by GWAS [11] including 7-DHC reductase (*DHCR7*), 25-hydroxylase (*CYP2R1*) and vitamin D-binding protein (*GC*). Additionally, some genes involved in vitamin D metabolism or vitamin D-related signal transduction pathways, including 24-hydroxylase (*CYP24A1*), 1,25-hydroxylase (*CYP27B1*) [12,13], vitamin D receptor (*VDR*) [14,15] and protein kinase C (PKC) family genes [16] have been identified as associated with MS risk or as modifying the vitamin D-MS relationship. Interestingly, most of the genes involved in vitamin D metabolism belong to the cytochrome p450 (CYP450) family, whose expression may be modulated by IFN-β [17,18]. Taken together, it is possible that some genes involved in vitamin D pathway may modulate the effects of IFN-β, by modulating the effect of IFN-β on 25(OH)D levels, or directly modulating its

effects on relapse risk. Identifying genetic factors which modulate IFN-β's effects could have import not merely in clarifying its mode of action, but could potentially lead to pharmacogenomic applications in clinical practice, whereby particular genotypes might be more responsive to IFN-β therapy.

To identify whether those genes involved in the vitamin D pathway modulate the effects of IFN-β on 25(OH)D and relapse, we utilised data from a prospective MS cohort study to evaluate whether differences exist in the IFN-β-25(OH)D relationship, the interaction between IFN-β and sun in predicting 25(OH)D, and the interaction between IFN-β and 25(OH)D in modulating relapse risk, at different levels of single nucleotide polymorphisms (SNPs) in genes relevant to vitamin D.

6.3 Methods

6.3.1 Study design

The Southern Tasmanian Multiple Sclerosis Longitudinal Study was designed as a prospective cohort study to evaluate the role of UV exposure and 25(OH)D on the clinical course of MS [7,19]. Briefly, this study followed a cohort of 198 persons with clinically definite MS (2001 McDonald criteria [20]) living in southern Tasmania, Australia between 2002 and 2005. Where participants discontinued participation or were lost to follow-up (4%; 8/198), they were censored at the date of study exit or their last attended review. Of the main cohort, 145 RRMS participants were followed beyond one review and of these 141 had data on genetic parameters under study. Ethics approval was obtained from the Southern Tasmania Human Research Ethics Committee. All participants provided informed consent.

Height (m) and weight (kg) were measured at study entry, and body mass index (BMI) was calculated as weight divided by height squared. Twenty persons with high disability (median Expanded Disability Status Scale [EDSS] 8.5, interquartile range [IQR] 8–9) could not have their height and weight measured and thus, BMI could not be calculated for these persons. Because BMI is an important determinant of 25(OH)D serum levels [21], these 20 people were excluded, leaving 178 persons. Of these, 169 participants with MS had genotype data.

At study entry and each biannual review, participants were asked about their lifestyle, including physical activity, smoking, fish intake, immunomodulatory therapy use, and vitamin D supplement use. Participants were queried for their average time in the sun on weekends/holidays in the two three-month periods prior to the date of review, with options including “never”, “less than half an hour a day”, “1/2 to 1 hour a day”, “1 to 2 hours a day”, “2 to 3 hours a day”, and “more than 3 hours a day”. Fish intake was queried by the Cancer Council of Victoria Food Frequency Questionnaire [22].

6.3.2 Measurement of 25(OH)D

These measurements have been described in detail elsewhere [5,7]. Briefly, at each biannual review, blood samples were taken. All samples were stored at -80°C and shielded from light. Serum 25(OH)D concentrations were measured using a commercially available radioimmunoassay (DiaSorin Inc., Stillwater, Minnesota). Assays were performed following the conclusion of the study, by a single operator in batches. Consequently, neither participants nor study personnel were aware of participants’ 25(OH)D concentrations during the study.

6.3.3 Genotyping

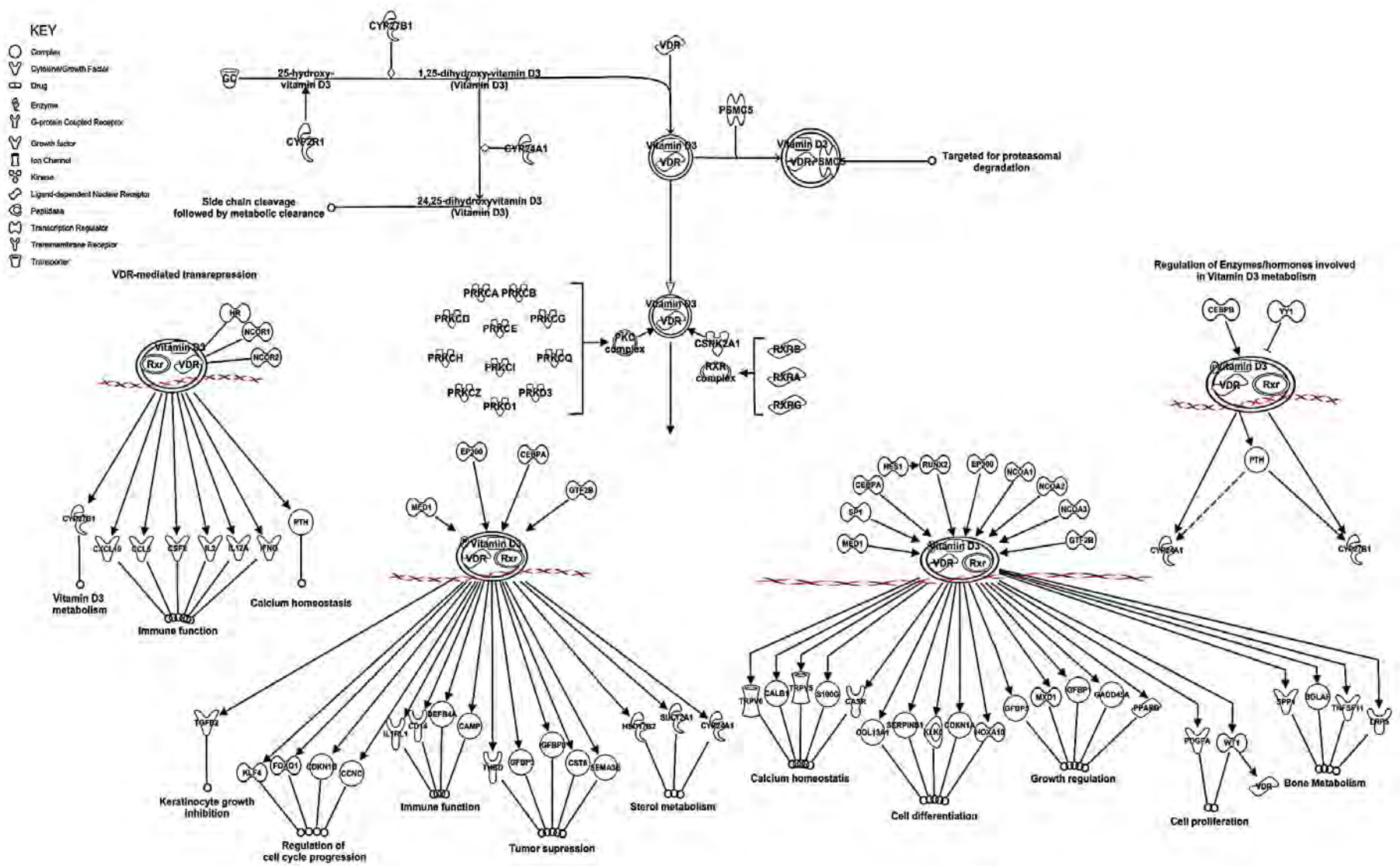
The methods have been described in detail in our previous study [16]. Briefly, 164 MS cases were genotyped as a part of ANZgene MS GWAS [12] and additionally 29 MS case were newly genotyped using the Illumina HumanOmniExpress-12v1_A array. All the samples were previously identified as being from individuals of European descent [12]. A conservative quality control was conducted with PLINK [23] before imputation: individuals with call rates less than 0.90, SNPs with call rates less than 0.95 or in Hardy-Weinberg ($p<10^{-7}$), or duplicates were excluded, leaving 189 cases with 290,536 SNPs.

6.3.4 Vitamin D pathway analysis and SNPs selection

As our *a priori* hypothesis was that genes involved in the vitamin D pathway may modulate the effects of IFN-β on serum vitamin D, we generated the canonical VDR/RXR pathway by Ingenuity Pathways Analysis (IPA; www.ingenuity.com), which is based on published interactions between gene products and the different forms of vitamin D. The vitamin D pathway is shown in **Figure 6. 1**. In total, there were 81 genes involved in the vitamin D pathway including six overlapping genes. For each of these genes we selected all SNPs from the Hg19 genome and 1000 Genome data [24] with minor allele frequency (MAF) ≥1% located within the physical boundaries, including 1kb upstream or downstream, and the genotyped SNPs for each of these genes were filtered further based on the genotype dataset. For those SNPs with high linkage disequilibrium (LD) in the same gene, TagSNPs were selected using the r^2-based tagger tool SNAP [25].

Figure 6. 1

The canonical VDR/RXR pathway. The module in the upside is the metabolic processes surrounding vitamin D and the formation of the transcription factor complex, and the downside is the various transcription factor complex pathways based on which co-factors bind the complex. This pathway was generated using the canonical pathway information stored in IPA (Ingenuity® Systems, www.ingenuity.com). Genes or gene products are represented as nodes, and the biological relationship between two nodes is represented as an edge (line). All relationships are supported by multiple published references. The shape of each node represents the functional class of the gene product, as shown in the key.


KEY
Complex
Cytokine/Growth Factor
Drug
Enzyme
G-protein Coupled Receptor
Growth factor
Ion Channel
Kinase
Ligand-dependent Nuclear Receptor
Peptidase
Transcription Regulator
Transmembrane Receptor
Transporter
CYP27B1
25-hydroxy-vitamin D3
1,25-dihydroxy-vitamin D3 (Vitamin D3)
VDR
PSMC5
CYP2R1
CYP24A1
Vitamin D3 VDR
Vitamin D3 VDR-PSMC5
Targeted for proteasomal degradation
Side chain cleavage followed by metabolic clearance
24,25-dihydroxyvitamin D3 (Vitamin D3)
VDR-mediated transrepression
PRKCA PRKCB
PRKCD
PRKCE
PRKCG
PRKCH
PRKCI
PRKCQ
PRKCZ
PRKD3
PRKD1
PKC complex
CSNK2A1
RXR complex
RXRB
RXRA
RXRG
HR
NCOR1
NCOR2
Rxr
Vitamin D3 metabolism
Immune function
Calcium homeostasis
PTH
EP300
CEBPA
MED1
GTF2B
P Vitamin D3 VDR Rxr
TGFB2
Keratinocyte growth inhibition
KLF4
FOXO1
CDKN1B
CCNC
Regulation of cell cycle progression
IL1RL1
CD14
DEFB4A
CAMP
Immune function
YWHD
GFBP3
GFBP1
CST6
SEMA3E
Tumor supression
HSD17B2
SLC27A1
Sterol metabolism
HES1
RUNX2
NCOA1
NCOA2
NCOA3
SP1
Vitamin D3 VDR Rxr
TRPV6
CALB1
TRPV5
S100G
CASR
Calcium homeostatis
IGFBP3
MXD1
GADD45A
PPARD
Growth regulation
SERPINB1
KLK6
CDKN1A
HOXA10
Cell differentiation
PDGFA
WT1
VDR
Cell proliferation
SPP1
BGLAP
TNFSF11
LRP5
Bone Metabolism
Regulation of Enzymes/hormones involved in Vitamin D3 metabolism
CEBPB
YY1
Vitamin D3 VDR Rxr
PTH
CYP24A1
CYP27B1

6.3.5 Data analysis

Predictors of 25(OH)D were evaluated by multilevel mixed-effects linear regression, with SNPs as the predictors and as-measured 25(OH)D as the outcome, adjusted for sun exposure, season, age, BMI, vitamin D supplementation, unexposed skin melanin density, weekly fish intake and smoking. Dependent variables were transformed as required to make the residuals less heteroskedastic; however all regression coefficients are reported on the scale of the original variable.

To assess the interaction between SNPs, sun exposure and IFN-β in predicting 25(OH)D, a double interaction model was employed. This model, using multilevel mixed-effects linear regression, made use of single interaction product terms between each of the pairs of the three primary covariates and a double interaction product term which included all three covariates. This double interaction term was evaluated to assess the significance of a difference in the IFN-β-sun interaction in predicting 25(OH)D by levels of the SNP under analysis.

The hazard of relapse was modelled as a function of 25(OH)D and other covariates on time-to-relapse and was calculated using Cox proportional hazards models for repeated events, using the gap-time model described by Prentice and colleagues [26]. Standard errors were adjusted to reflect multiple events per person. All covariates satisfied the proportional hazards assumption with the exception of the binary variable for sex and the categorical variable for baseline EDSS (0 – 2.5, 3.0 – 5.0, 5.5– 7.0, 7.5 – 9.0). For this reason, all models are stratified to allow the baseline hazards to differ by sex and baseline EDSS category. 25(OH)D was estimated at monthly intervals between biannual

measurements using methods described previously [7]. Briefly, a sinusoidal curve was fitted to the measured 25(OH)D and used to predict values at monthly intervals between measures. To examine whether there was an interaction between 25(OH)D and a SNP on relapse, a product term was included in the model, and then the association between 25(OH)D and hazard of relapse could be estimated for each SNP allele level, providing stratum-specific data.

Experiment-wide significance threshold for each gene was assessed using the tool of Single Nucleotide Polymorphism Spectral Decomposition (SNPSpD) [27], the effective number of independent marker loci (MeffLi) for each gene were calculated by SNPSpD. SNPs with a *p*-value after adjustment for covariates reached the experiment-wide significance threshold, or with the *p*-value less than 0.05 after adjustment for covariates and the MeffLi were considered to be significant.

All data analyses were performed using STATA/IC 12.1 (StataCorp LP, College Station, Texas, USA).

6.4 Results

The 169 participants who had BMI and genetic data comprised the analysis cohort. This group was followed for an average of 2.3 years, was 71.6% female and of mean age 47.8 (SD: 11.2 years), and mean baseline EDSS of 3.5. Of this group, 55.6% used IFN-β medication and 29.9% used vitamin D supplement during the study. The 141 participants who were of relapsing-remitting course and followed beyond one review, and for whom genotype data was obtained, comprised the cohort for relapse analyses, and had

demographics similar to the parent cohort. A total of 122 confirmed relapses occurred in 70 participants. Other characteristics of the cohort are shown in **Table 6. 1**.

Table 6.1. Demographic and clinical characteristics of study participants

	All persons, n (%)	RRMS, n (%)
Total	169	141
Sex		
Female	120 (71.6)	106 (75.1)
Male	49 (28.4)	35 (24.9)
Age at study entry, y		
21-39	41 (23.6)	38 (26.7)
39-45	36 (21.6)	34 (24.0)
45-52	39 (22.8)	33 (23.5)
52-76	53 (32.0)	36 (25.8)
BMI at study entry		
Normal	63 (37.5)	56 (40.5)
Overweight	70 (41.8)	54 (37.4)
Obese	36 (20.7)	31 (22.1)
Relapse during study?		
No		126 (89.4)
Yes		15 (10.6)
Ever smoked?		
No	64 (37.4)	59 (41.3)
Yes	105 (62.6)	82 (58.7)
Any vitamin D supplements during study?		
No	118 (70.1)	99 (70.2)
Yes	51 (29.9)	42 (29.8)
Any immunomodulatory therapy during study?		
No	47 (26.2)	25 (16.7)
Yes	122 (73.8)	116 (83.3)
Any interferon-β medications used during study?		
No	75 (44.4)	51 (36.2)
Yes	94 (55.6)	90 (63.8)
Age at study entry, y, mean (SD; range)	47.8 (11.2; 21-77)	45.9 (10.1; 21-76)
MS duration from diagnosis, y, median (IQR)	6 (2-12)	4 (2-10)
MS duration from 1st symptoms, y, median (IQR)	11.5 (6-19)	10 (5-17)
EDSS at study entry, median (IQR)	3 (2-4.5)	2.5(2-3.5)

Abbreviations: BMI= body mass index; EDSS= Expanded Disability Status Scale; IQR= interquartile range.

We found that two genotyped SNPs (rs10767935 and rs5030244) in the gene *WT1* significantly modified the relationship between IFN-β and 25(OH)D, such that IFN-β only increased 25(OH)D levels among those carrying the rare allele of rs10767935 and the major allele of rs5030244, these effects persisting even after adjustment for covariates and multiple comparisons for MeffLi ($p_{interaction}$=0.001, 0.0002; p_{adj}=0.003, 0.006,

respectively). *WT1* is a component of transcription factors complex 1 in the downstream of vitamin D pathway (**Figure 1**), and encodes a zinc finger DNA-binding protein that acts as a transcriptional activator or repressor. The experiment-wide significance threshold for *WT1* was 0.017, with the effective number of independent marker loci (MeffLi) of 3**.** Additionally, a clear allele dose-response was observed for these two SNPs, such that among subjects carrying at least one copy of the rare allele of rs10767935 (CT and TT genotype), those using IFN-β had significantly higher 25(OH)D by 13.32 nmol/L (p=4.19x10^{-7}), relative to those not using IFN-β medication. For rs5030244, subjects carrying major homozygous allele (TT genotype) with IFN-β use had significantly higher 25(OH)D by 9.05 nmol/L (p=8.32x10^{-8}) compared to those not using IFN-β medication. Among those carrying the homozygous major allele of rs10767935 (CC genotype) or at least one copy of minor allele of rs5030244, the levels of 25(OH)D did not differ by IFN-β use **(Table 6. 2).**

Table 6.2. Relationship of IFN-β and 25(OH)D by level of the significant SNPs in *WT1* in MS patients

Chr/gene	SNP	Genotype	IFN-β	N	25(OH)D (95% CI)	*p*
11/WT1	rs10767935	CC	none	48	46.38 (43.40, 49.35)	ref
			used	58	+2.54 (-1.09, 6.17)	*p=0.17*
		CT+TT	none	27	42.46 (38.79, 46.13)	ref
			used	36	+13.32 (8.16, 18.49)	p=4.19x10^{-7}
						$p_{interaction}$*=0.001*
						p_{adj}*=0.003*
11/WT1	rs5030244	CC+CT	none	12	52.52 (46.09, 58.95)	ref
			used	14	-5.18 (-12.02, 1.67)	*p=0.14*
		TT	none	63	43.34 (40.87, 45.80)	ref
			used	80	+9.05 (5.74, 12.36)	p=8.32x10^{-8}
						$p_{interaction}$ *=0.0002*
						p_{adj}*=0.0006*

$p_{interaction}$*:* adjusted for age, sex, sun exposure, BMI, smoking, baseline EDSS, season and vitamin D supplement use, before adjustment for multiple comparisons; P_{adj}*:* further adjusted for multiple comparisons.
The TT genotype of rs10767935 (n=7) and CC genotype of rs5030244 (n=2) were combined with the heterozygous genotype because of the low frequency.

We then estimated whether these two genotyped SNPs modified the relationship between IFN-β and 25(OH)D in the RRMS group. We found similar findings, these two SNPs significantly modifying the relationship between IFN-β and 25(OH)D, persisting after adjustment for covariates and multiple comparisons of MeffLi ($p_{interaction}$ =0.0003, 0.0004; p_{adj} =0.0009, 0.0012, respectively) (**Table 6. 3**). In addition, similar allele dose-responses were found for these two SNPs in the RRMS subgroup, with those individuals with IFN-β use and carrying at least one copy of the rare allele of rs10767935 or the major homozygous allele of rs5030244 had 25(OH)D levels significantly higher than those that didn't use IFN-β medication (**Table 6. 3**).

Table 6.3. Relationship of IFN-β and 25(OH)D by level of the significant SNPs in *WT1* in RRMS patients

Chr/gene	SNP	Genotype	IFN-β	N	25(OH)D (95% CI)	*p*
11/WT1	rs10767935	CC	none	31	48.57 (44.90, 52.24)	ref
			used	55	+2.00 (-2.06, 6.06)	*p=0.33*
		CT+TT	none	21	43.21 (38.99, 47.43)	ref
			used	34	+15.10 (9.58, 20.61)	$p=7.87x10^{-8}$
						$p_{interaction}=0.0003$
						$p_{adj}=0.0009$
11/WT1	rs5030244	CC+CT	none	11	52.61 (45.74, 59.47)	ref
			used	22	-5.08 (-12.32, 2.16)	*p=0.17*
		TT	none	41	44.75 (41.71, 47.79)	ref
			used	77	+9.63 (6.00, 13.26)	$p=1.89x10^{-7}$
						$p_{interaction}=0.0004$
						$p_{adj}=0.0012$

$p_{interaction}$*:* adjusted for age, sex, sun exposure, BMI, smoking, baseline EDSS, season and vitamin D supplement use, before adjustment for multiple comparisons; P_{adj}*:* further adjusted for multiple comparisons. The TT genotype of rs10767935 (n=6) and CC genotype of rs5030244 (n=2) were combined with the heterozygous genotype because of the low frequency.

Previously, we found that IFN-β therapy significantly potentiated the production of 25(OH)D from sun exposure, such that those on therapy realised nearly three-times greater 25(OH)D from equivalent sun exposure [5]. We next examined whether these two SNPs in *WT1* modulated the association between sun exposure and IFN-β in predicting

25(OH)D levels. A notable trend was evident, showing the interactive relationship between IFN-β and sun exposure in predicting 25(OH)D differed appreciably by level of the two SNPs, such that those carrying at least one copy of the T allele (CT and TT) of rs10767935 or those homozygous for the T allele of rs5030244 demonstrated a marked difference in the sun-25(OH)D association by IFN-β therapy, whereas those homozygous for the C allele of rs10767935 or carrying any C allele (CC and CT) of rs5030244 showed no difference. These differences did not reach statistical significance, however (**Table 6.4 -6.5; Figure 6.2 -6.3**). Analogous interactions were seen for the associations of fish intake and vitamin D supplement dose in predicting 25(OH)D, but these too did not reach statistical significance (**Table 6.6-6.7**). No interactions with sun were found for the other *WT1* SNP, rs5030244, however (**Table 6.8)**, and trends for fish intake and vitamin D supplement use were less clear than those seen for rs10767934 (**Tables 6.9-6.10**).

Table 6.4. Interactive relationship between sun and IFN-β in predicting 25(OH)D, by level of SNP rs10767935 in *WT1*

	Mean 25(OH)D (95% CI)		
	Not using IFN-β	Using IFN-β	Sun-IFN-β interaction
All persons			
<1/2	44.86 (41.47, 48.26)	45.84 (42.50, 49.19)	
1/2 – <1	-3.20 (-6.84, 0.44)	+2.49 (-1.35, 6.33)	
1 – <2	-1.80 (-6.12, 2.51)	+4.70 (0.73, 8.67)	
2 - <3	+6.01 (1.06, 10.96)	+ 12.05 (6.87, 17.22)	
3 +	+3.90 (-1.38, 9.19)	+ 19.26 (12.69, 25.83)	
Trend:	$p=1.12x10^{-6}$	$p=1.13x10^{-22}$	*p=0.01*
rs10767935=CC			
<1/2	42.56 (38.59, 46.54)	43.31 (39.81, 46.81)	
1/2 – <1	+2.05 (0.18, 3.92)	+3.45 (2.10, 4.80)	
1 – <2	+4.20 (0.26, 8.14)	+7.18 (4.24, 10.11)	
2 - <3	+6.46 (0.25, 12.66)	+ 11.21 (6.42, 15.99)	
3 +	+8.82 (0.13, 17.52)	+ 15.56 (8.64, 22.48)	
Trend:	*p*=0.026	$p= 1.30x10^{-4}$	*p=0.20*
rs10767935=CT			
<1/2	42.20 (37.22, 47.17)	47.20 (41.88, 52.53)	
1/2 – <1	+0.85 (-1.28, 2.98)	+4.61 (3.40, 5.81)	
1 – <2	+1.72 (-2.62, 6.06)	+ 9.67 (6.99, 12.34)	
2 - <3	+2.60 (-4.05, 9.25)	+ 15.23 (10.78, 19.68)	
3 +	+3.50 (-5.54, 12.55)	+ 21.34 (14.76, 27.92)	
Trend:	*p=0.40*	$p= 3.60x10^{-5}$	*p=0.11*
rs10767935=TT			
<1/2	41.83 (31.53, 52.14)	51.45 (39.57, 63.33)	
1/2 – <1	-0.30 (-4.81, 4.21)	+5.96 (3.39, 8.53)	
1 – <2	-0.60 (-9.54, 8.35)	+ 12.62 (6.80, 18.44)	
2 - <3	-0.89 (-14.22, 12.43)	+ 20.06 (10.19, 29.94)	
3 +	-1.18 (-18.82, 16.45)	+ 28.38 (13.50, 43.27)	
Trend:	*p=0.72*	*p*=0.23	*p=0.12*
Sun-IFN-β-SNP interaction	*p=0.23*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.

Table 6.5. Interactive relationship between sun and IFN-β in predicting 25(OH)D, overall and by level of SNP rs5030244 in *WT1*

	Mean 25(OH)D (95% CI)		
	Not using IFN-β	Using IFN-β	Sun-IFN-β interaction
All persons			
<1/2	44.78 (41.39, 48.18)	45.80 (42.46, 49.14)	
1/2 – <1	-3.18(-6.81, 0.46)	+2.52 (-1.32, 6.36)	
1 – <2	-1.70 (-6.02, 2.61)	+4.75 (0.78, 8.72)	
2 - <3	+5.99 (1.05, 10.93)	+12.12 (6.94, 17.31)	
3 +	+4.00 (-1.28, 9.29)	+19.40 (12.82, 25.98)	
Trend:	$p=1.12x10^{-6}$	$p=1.13x10^{-22}$	*p=0.01*
rs5030244 = CC+CT			
<1/2	46.15 (40.26, 52.04)	39.40 (34.01, 44.79)	
1/2 – <1	+3.81 (0.35, 7.27)	+4.87 (2.17, 7.57)	
1 – <2	+7.94 (0.43, 15.46)	+10.35 (4.23, 16.48)	
2 - <3	+12.43 (0.18, 24.68)	+16.54 (6.13, 26.94)	
3 +	+17.29 (-0.46, 35.05)	+23.51 (7.78, 39.23)	
Trend:	*p=0.01*	$p=0.004$	*p=0.51*
rs5030244 =TT			
<1/2	41.40 (37.75, 45.05)	46.06 (42.48, 49.64)	
1/2 – <1	+1.19 (-0.44, 2.82)	+3.80 (2.66, 4.94)	
1 – <2	+2.42 (-0.95, 5.78)	+7.92 (5.43, 10.40)	
2 - <3	+ 3.68 (-1.52, 8.87)	+12.38 (8.32, 16.45)	
3 +	+ 4.98 (-2.16, 12.11)	+17.23 (11.32, 23.14)	
Trend:	$p=0.09$	$p=5.60x10^{-7}$	*p=0.033*
Sun-IFN-β-SNP interaction	*p=0.81*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.
The CC genotype was combined with CT genotype because of the low frequency (n=2).

Table 6.6. Interactive relationship between average weekly fish intake and IFN-β in predicting 25(OH)D, overall and by level of SNP rs10767935 in *WT1*

	Mean 25(OH)D (95% CI)		**Sun-IFN-β**
	Not using IFN-β	**Using IFN-β**	**interaction**
All persons			
<1 servings/wk	44.30 (40.19, 48.41)	51.23 (47.48, 54.98)	
1 – 2 servings/wk	+0.98 (-3.91, 5.87)	+0.72 (-3.20, 4.64)	
2 – 4 servings/wk	-1.29 (-6.52, 3.95)	+3.09 (-1.46, 7.63)	
>4 servings/wk	+0.54 (-5.12, 6.20)	+4.02 (-1.45, 9.48)	
Trend:	*p*=0.86	*p*=0.12	*p=0.16*
rs10767935=CC			
<1 servings/wk	54.78 (49.72, 59.85)	58.36 (53.08, 63.65)	
1 – 2 servings/wk	-0.84 (-3.60, 1.92)	+2.00 (-0.57, 4.57)	
2 – 4 servings/wk	-1.67 (-7.10, 3.77)	+4.07 (-1.25, 9.39)	
>4 servings/wk	-2.48 (-10.51, 5.55)	+6.21 (-2.05, 14.47)	
Trend:	*p*=0.47	*p*=0.22	*p=0.13*
rs10767935=CT			
<1 servings/wk	52.03 (46.79, 57.26)	65.31 (58.70, 71.91)	
1 – 2 servings/wk	+0.84 (-1.75, 3.42)	+1.19 (-2.60, 4.97)	
2 – 4 servings/wk	+1.69 (-3.57, 6.94)	+2.39 (-5.32, 10.11)	
>4 servings/wk	+2.55 (-5.46, 10.56)	+3.63 (-8.16, 15.41)	
Trend:	*p*=0.47	*p*=0.16	*p=0.78*
rs10767935=TT			
<1 servings/wk	49.42 (41.00, 57.84)	73.09 (59.78, 86.41)	
1 – 2 servings/wk	+2.39 (-2.95, 7.73)	+0.17 (-8.44, 8.77)	
2 – 4 servings/wk	+4.91 (-6.31, 16.12)	+0.33 (-16.91, 17.57)	
>4 servings/wk	+7.54 (-10.12, 25.20)	+0.50 (-25.43, 26.42)	
Trend:	*p*=0.44	*p*=0.64	*p=0.38*
Sun-IFN-β-SNP interaction	*p=0.30*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.

Table 6.7. Interactive relationship between vitamin D supplement use and IFN-β in predicting 25(OH)D, overall and by level of SNP rs10767935 in *WT1*

	Mean 25(OH)D (95% CI)		Sun-IFN-β
	Not using IFN-β	Using IFN-β	interaction
All persons			
No	43.16 (40.76, 45.55)	51.39 (48.36, 54.43)	
Yes	+2.55 (-1.03, 6.12)	+2.31 (-1.20, 5.82)	
Trend:	*p=0.16*	*p=0.20*	*p=0.87*
rs10767935 = CC			
No	48.56 (43.70, 53.42)	51.02, 45.96, 56.07)	
Yes	+4.14 (-2.33, 10.60)	+2.89 (-2.36, 8.13)	
Trend:	*p=0.21*	*p=0.28*	*p=0.42*
rs10767935 = CT			
No	44.18 (37.31, 51.06)	56.49 (48.63, 64.34)	
Yes	+0.81 (-6.07, 7.69)	+2.17 (-6.14, 10.48)	
Trend:	*p=0.82*	*p=0.61*	***p=0.006***
rs10767935 = TT			
No	40.20 (26.99, 53.42)	62.56 (44.24, 80.87)	
Yes	-1.77 (-13.51, 9.97)	+1.29 (-17.23, 19.80)	
Trend:	*p=0.77*	*p=0.89*	***p<0.001***
Sun-IFN-β-SNP interaction	*p=0.74*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.

Table 6.8. Interactive relationship between average weekly fish intake and IFN-β in predicting 25(OH)D, overall and by level of SNP rs5030244 in *WT1*

	Mean 25(OH)D (95% CI)		**Sun-IFN-β**
	Not using IFN-β	**Using IFN-β**	**interaction**
All persons			
<1 servings/wk	44.30 (40.19, 48.41)	51.23 (47.48, 54.98)	
1 – 2 servings/wk	+0.98 (-3.91, 5.87)	+0.72 (-3.20, 4.64)	
2 – 4 servings/wk	-1.29 (-6.52, 3.95)	+3.09 (-1.46, 7.63)	
>4 servings/wk	+0.54 (-5.12, 6.20)	+4.02 (-1.45, 9.48)	
Trend:	*p*=0.86	*p*=0.12	*p=0.16*
rs5030244=CC+CT			
<1 servings/wk	63.90 (53.88, 73.91)	57.78 (48.90, 66.67)	
1 – 2 servings/wk	-0.91 (-7.24, 5.43)	-3.26 (-8.08, 1.57)	
2 – 4 servings/wk	-1.80 (-14.27, 10.68)	-6.33 (-15.43, 2.78)	
>4 servings/wk	-2.68 (-21.12, 15.77)	-9.22 (-22.10, 3.67)	
Trend:	*p*=0.52	*p*=0.86	*p=0.60*
rs5030244=TT			
<1 servings/wk	51.75 (47.52, 55.98)	62.09 (56.76, 67.42)	
1 – 2 servings/wk	+0.29 (-2.00, 2.58)	+2.63 (0.16, 5.10)	
2 – 4 servings/wk	+0.58 (-4.03, 5.20)	+5.38 (0.22, 10.54)	
>4 servings/wk	+0.88 (-6.08, 7.84)	+8.25 (0.17, 16.33)	
Trend:	*p*=0.66	*p*=0.044	*p=0.21*
Sun-IFN-β-SNP interaction	*p=0.27*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.
The CC genotype was combined with CT genotype because of the low frequency (n=2).

Table 6.9. Interactive relationship between vitamin D supplement use and IFN-β in predicting 25(OH)D, overall and by level of SNP rs5030244 in *WT1*

	Mean 25(OH)D (95% CI)		Sun-IFN-β
	Not using IFN-β	Using IFN-β	interaction
All persons			
No	43.16 (40.76, 45.55)	51.39 (48.36, 54.43)	
Yes	+2.55 (-1.03, 6.12)	+2.31 (-1.20, 5.82)	
Trend:	*p=0.16*	*p=0.20*	*p=0.87*
rs5030244=CC+CT			
No	60.30 (43.19, 77.41)	42.68 (32.39, 52.97)	
Yes	+12.26 (-11.97, 36.48)	-4.45 (-12.29, 3.39)	
Trend:	*p=0.32*	*p=0.27*	*p=0.13*
rs5030244=TT			
No	45.02 (41.28, 48.76)	55.22 (50.44, 60.01)	
Yes	+2.44 (-2.34, 7.21)	+4.34 (-0.97, 9.65)	
Trend:	*p=0.32*	*p=0.11*	*p=0.77*
Sun-IFN-β-SNP interaction	*p=0.11*		

Geometric mean 25(OH)D and coefficients relative to reference mean by level of sun exposure and IFN-β adjusted for age, sex, smoking, BMI, season, baseline EDSS and vitamin D supplement use.
The CC genotype was combined with CT genotype because of the low frequency (n=2).

Figure 6.2

Interactive relationship between sun and IFN-β in predicting 25(OH)D, overall and by level of SNP rs10767935 in *WT1*.

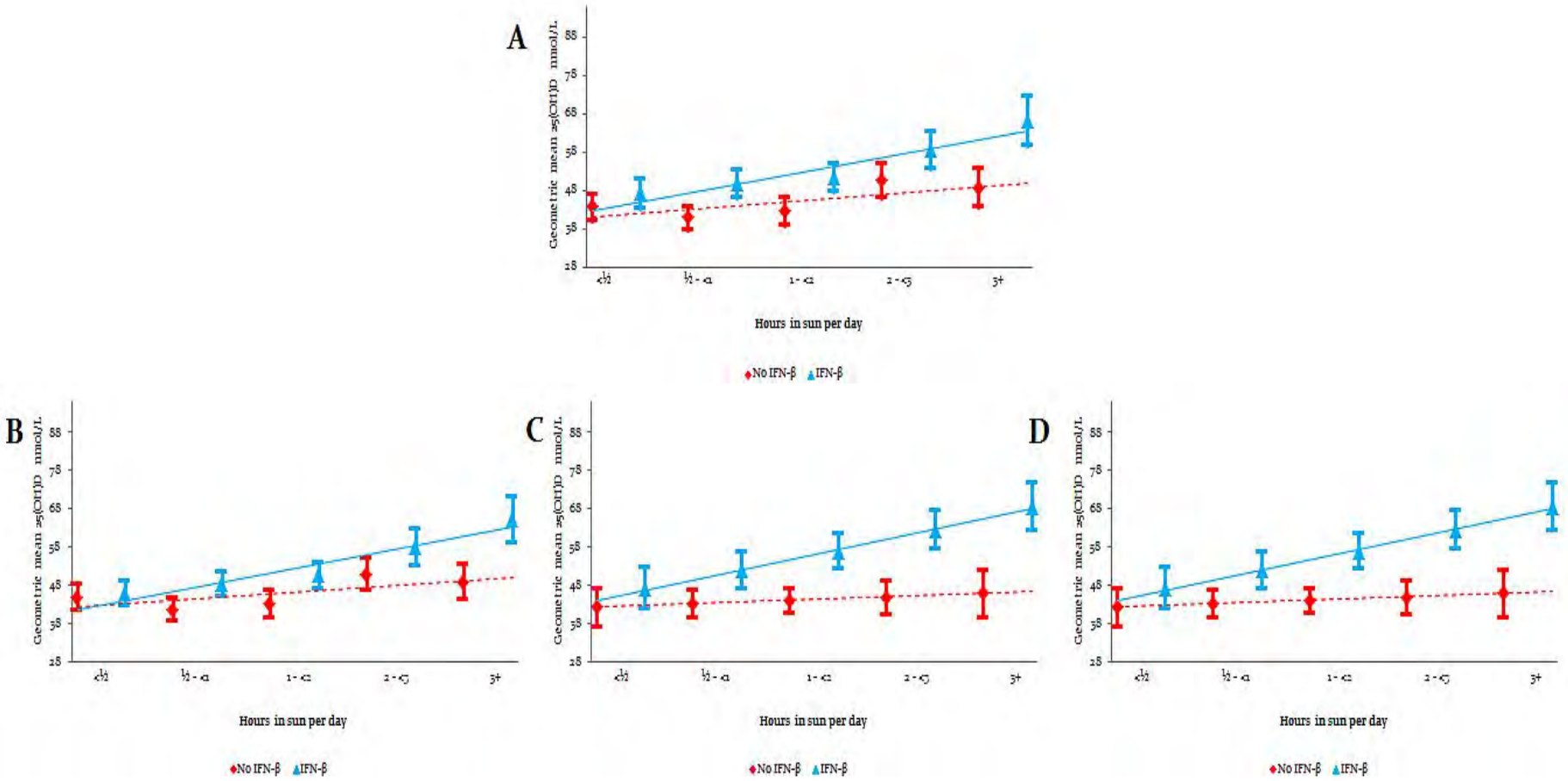
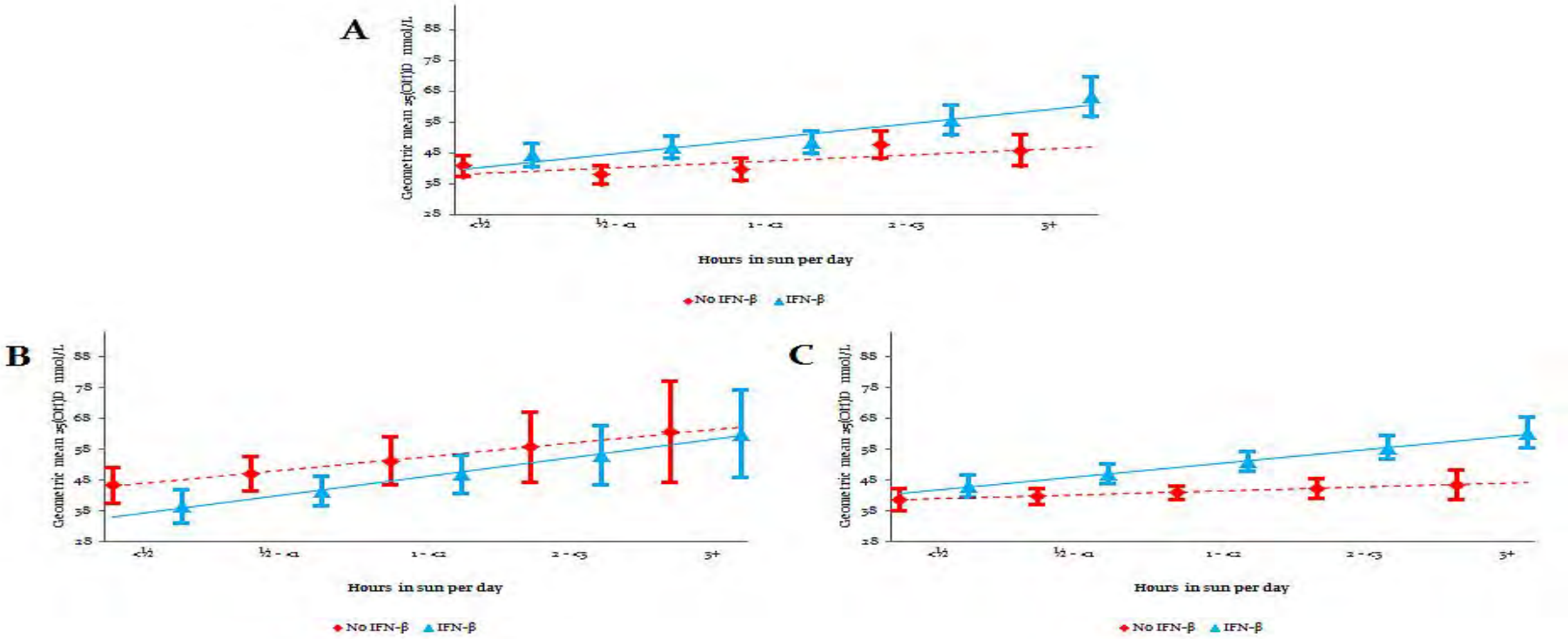


Figure 6.3

Interactive relationship between sun and IFN-β in predicting 25(OH)D, overall and by level of SNP rs5030244 in *WT1*.



We next evaluated whether these two SNPs in *WT1* significantly impacted the interaction between IFN-β and 25(OH)D in predicting relapse. We previously demonstrated that the association of 25(OH)D with relapse significantly differed by IFN-β therapy, such that 25(OH)D was only inversely associated with relapse among persons on therapy [5]. When we evaluated whether this interaction differed by level of the two SNPs, we found no evidence of a difference in this interaction by level of either rs10767935 or rs5030244 (**Table 6.10**)

Table 6.10. Interactive relationship between IFN-β & 25(OH)D in predicting relapse risk, overall and by level of SNP

Chr/gene	SNP	Genotype	IFN-β	N	HR of 25(OH)D (95% CI)	*p*	IFN-β & 25(OH)D interaction
11/WT1	rs10767935	All persons	none	48	0.97 (0.86, 1.09)	*p=0.57*	ref
			used	93	0.80 (0.71, 0.89)	*p<0.001*	*p=0.001*
		CC	none	31	1.15 (0.60, 2.22)	*p=0.67*	ref
			used	55	0.59 (0.33, 1.06)	*p=0.078*	*p=0.018*
		CT+TT	none	21	2.42 (0.97, 6.05)	*p=0.058*	*ref*
			used	34	0.25 (0.11, 0.55)	*p=0.001*	*p=0.009*
							$p_{interaction}$ *=0.58*
11/WT1	rs5030244	All persons	none	48	0.97 (0.86, 1.09)	*p=0.57*	ref
			used	93	0.80 (0.71, 0.89)	*p<0.001*	*p=0.001*
		CC+CT	none	11	1.98 (0.30, 13.22)	*p=0.48*	ref
			used	12	0.30 (0.09, 1.06)	*p=0.062*	*p=0.091*
		TT	none	41	1.64 (0.89, 3.02)	*p=0.11*	ref
			used	77	0.48 (0.29, 0.80)	*p=0.005*	*p=0.004*
							$p_{interaction}$ *=0.57*

$p_{interaction}$*:* adjusted for age, sex, and baseline EDSS.
The TT genotype of rs10767935 and CC genotype of rs5030244 were combined with the heterozygous genotype because of the low frequency (n=6 and n=2, respectively).

6.5 Discussion

In this prospective cohort study, we found that two genotype SNPs (rs10767935 and rs5030244) in *WT1* significantly modified the previously demonstrated association

between IFN-β and 25(OH)D in patients with MS and RRMS, persisting even after adjustment for relevant covariates and for multiple comparisons.

These two SNPs locate in the introns of the Wilms tumor 1 gene, which encodes a zinc finger DNA-binding protein that acts as a transcriptional activator or repressor, which has been demonstrated to modulate the response to active form of vitamin D (1,25-dihydroxyvitamin D3) by induction of VDR [28]. In addition, the human VDR gene has been confirmed as a downstream target of WT1, with three *WT1*-responsive sites identified in the *VDR* promoter [29]. Interestingly, *WT1* is involved in cell proliferation as a co-factor with *PDGFA* gene and also binds to VDR in the downstream of vitamin D pathway (**Figure 6.1**). These findings suggested the effect of *WT1* variants may be mediated by VDR activation in downstream signalling pathways which consequently influence 25(OH)D levels. However, the mechanism how 25(OH)D level is impacted by the transcriptional activation by *WT1* on VDR is unclear.

These two SNPs in *WT1* significantly modulated the relationship between IFN-β and 25(OH)D, with those individuals on IFN-β therapy and carrying at least one copy of the rare allele of rs10767935 or the major homozygous allele of rs5030244 having significantly higher 25(OH)D levels than those not using IFN-β medication. Notably, we reported previously that IFN-β treatment was associated with higher 25(OH)D among persons with MS apparently mediated by a stronger association between personal sun exposure and subsequently measured 25(OH)D[5]. In the present analysis, we showed a marked difference in the interaction between IFN-β use and sun in predicting 25(OH)D by level of the rs10767935 SNP, such that carriers of the major allele of rs10767935 (CC) showed no significant difference in the sun-25(OH)D relationship between those on or off

IFN-β treatment, while those carrying the minor allele (CC/TT) showed a five-fold greater sun-25(OH)D link among those on IFN-β therapy compared to those not on therapy. Despite the marked differences in the interaction between IFN-β and sun predicting by level of rs10767935, this interaction did not reach statistical significance, likely due in part to the small numbers in our sample and the double interaction method employed which reduced statistical power. Despite the significant interaction with IFN-β in predicting 25(OH)D, no interaction with IFN-β and sun in predicting 25(OH)D was evident for the rs5030244 SNP. That we did not see a similar interaction with sun for the rs5030244 SNP despite the significant difference in the IFN-β effect by level of the SNP reflect sample size and methodological limitations, or it may indicate this SNP acts on vitamin D by a different pathway than sun exposure.

A core strength of our study was the availability of detailed environmental data, including potential confounders such as vitamin D supplementation and fish intake. A weakness of our study is the sample size. Although this is one of the largest and most well-studied MS cohorts available, the limitation imposed by the sample size reflects the difficulty of undertaking studies of this nature. However, we have used other methodologies to overcome potential type 1 error, including allele dose responses and the consistency of associations between multiple measures of association, namely direct prediction of 25(OH)D and interaction with sun in predicting 25(OH)D. These findings provide further support for a real association for the significant SNPs.

In conclusion, this is the first prospective study to demonstrate two SNPs (rs10767935 and rs5030244) in *WT1* significantly modified the previously demonstrated association between IFN-β and 25(OH)D in the patients with MS. While replication studies and

investigations into the molecular mechanism by how *WT1* alter therapeutic effects of IFN-β are required, these findings indicate that *WT1* variants may play an important role in the altering therapeutic effects of IFN-β. These results may suggest *WT1* variants may play some role in the function of IFN-β and may potentially have application in pharmacogenomic-informed allocation of IFN-β in clinical practice.

6.6 Summary

Objective: To investigate whether those genes involved in the vitamin D pathway modulate the demonstrated relationship between 25(OH)D and IFN-β, the relationship between IFN-β and sun in predicting 25(OH)D, and the interaction between IFN-β and 25(OH)D in modulating relapse risk in people with MS.

Methods: We conducted a prospective cohort study of 169 participants with MS and genotype data followed from 2002 to 2005. Gene-IFN-β and gene-IFN-β-sun interactions predicting 25(OH)D were evaluated by multilevel mixed-effects linear regression. Gene-IFN-β interactions with 25(OH)D in modulating relapse risk were assessed using survival analysis. Experiment-wide significance threshold was adjusted by the effective number of independent marker loci (MeffLi) for each gene.

Results: Two intronic genotyped SNPs (rs10767935 and rs5030244) in the gene *WT1* significantly modified the relationship between IFN-β and 25(OH)D even after adjustment for covariates and multiple comparisons for MeffLi in MS patients ($p_{interaction}$=0.001, 0.0002; p_{adj}=0.003, 0.006, respectively) and in RRMS subgroup ($p_{interaction}$ =0.0003, 0.0004; p_{adj} =0.0009, 0.0012, respectively). There was a marked

difference in the interaction between sun and IFN-β in predicting 25(OH)D by level of the rs10767935 *WT1*SNP, though this did not reach statistical significance. There were no differences in the interaction between 25(OH)D and IFN-β in predicting relapse by level of either *WT1* SNP.

Conclusions: We have demonstrated for the first time that two SNPs (rs10767935 and rs5030244) in *WT1* significantly modified the previously demonstrated association between IFN-β and 25(OH)D in the patients with MS. Some evidence was shown for a difference in the sun-IFN-β-25(OH)D association by level of rs10767935, but no difference by SNP was found for the 25(OH)D-IFN-β-relapse association. These findings indicate that *WT1* variants may play a role in altering the therapeutic effects of IFN-β in MS.

6.7 Postscript

This chapter has discussed how to detect the interactions between IFN-β and the genes involved in the vitamin D pathway influencing the levels of serum vitamin D. We demonstrated that the *WT1* gene significantly modified the association between IFN-β and 25(OH)D in patients with MS and RRMS, and showed evidence for modulating the sun-IFN-β-25(OH)D association and the 25(OH)D-IFN-β-relapse association. These findings indicate that *WT1* may play an important role in altering the therapeutic effects of IFN-β in MS. In the next chapter I will make my final conclusions from my PhD research and discuss the potential impact of my PhD research output, and will present my own views on the future direction of MS research.

6.8 References

1. PRISMS Study Group and the University of British Columbia MS/MRI Analysis Group. (2001) PRISMS-4: Long-term efficacy of interferon-beta-1a in relapsing MS. Neurology 56: 1628-1636.
2. Noronha A, Toscas A, Jensen MA (1993) Interferon beta decreases T cell activation and interferon gamma production in multiple sclerosis. J Neuroimmunol 46: 145-153.
3. Markowitz CE (2007) Interferon-beta: mechanism of action and dosing issues. Neurology 68: S8-11.
4. Yong VW, Chabot S, Stuve O, Williams G (1998) Interferon beta in the treatment of multiple sclerosis: mechanisms of action. Neurology 51: 682-689.
5. Stewart N, Simpson SJ, van der Mei I, Ponsonby AL, Blizzard L, et al. (2012) Interferon-β and serum 25-hydroxyvitamin D interact to modulate relapse risk in MS. Neurology 79: 254-260.
6. Ascherio A, Munger KL, Simon KC (2010) Vitamin D and multiple sclerosis. Lancet Neurology 9: 599-612.
7. Simpson S, Jr., Taylor B, Blizzard L, Ponsonby AL, Pittas F, et al. (2010) Higher 25-hydroxyvitamin D is associated with lower relapse risk in multiple sclerosis. Ann Neurol 68: 193-203.
8. Runia TF, Hop WC, de Rijke YB, Buljevac D, Hintzen RQ (2012) Lower serum vitamin D levels are associated with a higher relapse risk in multiple sclerosis. Neurology 79: 261-266.
9. Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A (2006) Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. Jama 296: 2832-2838.
10. Mowry E, Krupp L, Milazzo M, Chabas D, Strober J, et al. (2010) Vitamin D status is associated with relapse rate in pediatric-onset MS. Annals of Neurology 67: 618-624.
11. Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, et al. (2010) Common genetic determinants of vitamin D insufficiency: a genome-wide association study. Lancet 376: 180-188.
12. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet 41: 824-828.
13. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
14. Simon KC, Munger KL, Xing Y, Ascherio A (2010) Polymorphisms in vitamin D metabolism related genes and risk of multiple sclerosis. Mult Scler 16: 133-138.
15. Dickinson JL, Perera DI, van der Mei AF, Ponsonby AL, Polanowski AM, et al. (2009) Past environmental sun exposure and risk of multiple sclerosis: a role for the Cdx-2 Vitamin D receptor variant in this interaction. Mult Scler 15: 563-570.
16. Lin R, Taylor BV, Simpson SJ, Charlesworth J, Stankovich J, et al. (2013) Novel modulating effects of PKC family genes on the relationship between serum vitamin D and relapse in multiple sclerosis. J Neurol Neurosurg Psychiatry, in press.
17. Omdahl JL, Morris HA, May BK (2002) Hydroxylase enzymes of the vitamin D pathway: expression, function, and regulation. Annu Rev Nutr 22: 139-166.

18. Vidal M, Ramana CV, Dusso AS (2002) Stat1-vitamin D receptor interactions antagonize 1,25-dihydroxyvitamin D transcriptional activity and enhance stat1-mediated transcription. Mol Cell Biol 22: 2777-2787.
19. Pittas F, Ponsonby AL, van der Mei IA, Taylor BV, Blizzard L, et al. (2009) Smoking is associated with progressive disease course and increased progression in clinical disability in a prospective cohort of people with multiple sclerosis. J Neurol 256: 577-585.
20. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, et al. (2001) Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 50: 121-127.
21. Konradsen S, Ag H, Lindberg F, Hexeberg S, Jorde R (2008) Serum 1,25-dihydroxy vitamin D is inversely associated with body mass index. Eur J Nutr 47: 87-91.
22. Hodge A, Patterson AJ, Brown WJ, Ireland P, Giles G (2000) The Anti Cancer Council of Victoria FFQ: relative validity of nutrient intakes compared with weighed food records in young to middle-aged women in a study of iron supplementation. Aust N Z J Public Health 24: 576-583.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.
24. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, et al. (2012) The 1000 Genomes Project: data management and community access. Nature Methods 9: 459-462.
25. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938-2939.
26. Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. Biometrika 68: 373-379.
27. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74: 765-769.
28. Maurer U, Jehan F, Englert C, Hubinger G, Weidmann E, et al. (2001) The Wilms' tumor gene product (WT1) modulates the response to 1,25-dihydroxyvitamin D3 by induction of the vitamin D receptor. J Biol Chem 276: 3727-3732.
29. Lee TH, Pelletier J (2001) Functional characterization of WT1 binding sites within the human vitamin D receptor gene promoter. Physiol Genomics 7: 187-200.

Chapter 7. Conclusion

7.1 Rare variants/disease-causing variants and MS

The first two chapters (Chapter 2 and Chapter 3) focused on detecting rare disease-causing genetic variants that confer susceptibility to MS by using two approaches: identity-by-decent (IBD) mapping and family-based whole genome sequencing analysis. By utilising IBD mapping in a large GWAS dataset of 3,543 cases and 5,898 controls, we identified a genome-wide significant linkage signal on chromosome 19 (LOD=4.65; $p=1.9\times10^{-6}$). This linkage region includes a cluster of zinc finger protein genes of unknown function. Research into zinc finger genes demonstrated that some of them acting as transcript factors are involved in signaling pathways [1], regulation of cell proliferation and migration [2-7], cell survival [8] and DNA methylation [5,9,10], consequently influencing the development of a variety of types of cancer, including gallbladder cancer [2], gastric cancer [5,11-13], breast cancer [9], prostate cancer [9,14], colorectal cancer [15,16], oropharyngeal cancer [17], hepatocellular carcinoma [18] and type 1 diabetes (T1D) [3].

Little is known about zinc finger genes that are associated with MS, however, and none of the identified clusters of zinc finger genes have been previously identified in published GWAS or associated with MS or autoimmune diseases. When we examined the putative roles in CNS function for the genes in the linkage region in public human expression datasets, particularly the expression in foetal cells, we found most of the genes in this region exhibited high hippocampal expression in the foetus with low postnatal and adult expression, which suggested most genes in this linkage region are involved in early

developmental regulation of the CNS. Hence our hypothesis that a defect in CNS function is laid down early in development, which leads to MS susceptibility.

Unfortunately, the identified genome-wide significant linkage signal lies in a region that had many gene duplications and is hard to validate via sequencing, thus replication in other independent datasets is needed. Firstly using family-based sequencing data is an option. In family data, the common ancestor is fairly recent with long IBD segments, which is easier to detect even with low density genotype data, and with lower false discovery rates [19]. Alternatively, using population-based next-generation sequencing data is another powerful way. Since the IBD segments between two putatively unrelated subjects in a large population is fairly small, deeper coverage of genetic variants obtained through sequencing can increase the power to detect small IBD segments [20]. Exome sequencing, an assay of choice to efficiently sequence the coding regions of the genome, has enabled efficient discovery of highly penetrant variants that are too rare to be detected by SNP-array studies [21]. Early findings from family-based exome studies have demonstrated the power of using IBD segments to localize potential causal coding variants [22]. However, exome sequencing contains highly irregular density of read targets, and thus poses problems of accuracy [23]. Ideally, whole-genome sequencing is a means of providing higher quality data with less variation in site density. Recent studies have applied IBD detection algorithms to whole-genome sequences and yielded very high detection accuracy [20]. Taken together, it may be promising to perform IBD on merged exome sequencing and genotype data in conjunction, or using whole-genome sequencing data alone. To date the number of MS whole-genome sequences available is very small.

In Chapter 3 we conducted whole-genome sequencing analysis to assess disease-causing

variants to MS in a family that comprised three affected and four unaffected siblings and a daughter of a deceased sibling with known MS. We identified nine candidate variants by whole genome sequencing with further validation by Sanger sequencing. Amongst the nine candidate genes, two novel coding variants (chr6_31921581 and chr6_51930780) located in *RDBP* and *PKHD1* suggested to be promising potential disease-causing variants to MS.

RDBP is part of a complex termed negative elongation factor (NELF) that represses RNA polymerase II transcript elongation. *RDBP* localises to the major histocompatibility complex (MHC) class III region on chromosome 6, which has been demonstrated to be associated with MS in a GWAS [24]. Furthermore, the GERP score, Polyphen2 and SIFT showed evidence that the variant of chr6_31921581 has an impact on protein function. Interestingly, further test in our combined MS GWAS dataset also support that chr6_31921581 locates in a region that genome-wide significantly associated with MS, and significantly associated with MS in the Tasmanian group. As for *PKHD1,* which has been demonstrated as associated with ALS [25], is predicted to have a single transmembrane (TM)-spanning domain and multiple copies of an immunoglobulin-like plexin transcription-factor domain. We found that the variant of chr6_51930780 was dominant in the affected siblings, and Polyphen2 and SIFT also support its function to impact protein damage, suggesting chr6_51930780 is likely to be a disease-causing variant to MS.

This chapter presents our preliminary findings, and future analyses of the data are outside the limits of this thesis. Our findings have demonstrated the candidate disease-causing variants to MS, which have build up the foundation for the further research, such as

validation with our exome array data and replication in another MS family in Tasmania.

7.2 Gene-vitamin D interactions and MS

Chapter 4 and Chapter 5 focused on detecting gene-environment interactions that were associated with MS clinical course using a prospective cohort design, particularly focusing on the interactions between serum vitamin D metabolites and known MS-risk susceptibility variants or those in genes involved in the vitamin D pathway. Although we didn't identify any known MS susceptibility variants associated with relapse after adjustment for multiple testing, we did find that five SNPs were associated with relapse (p_{trend} <0.05), with significant cumulative genotype risk effects, and two demonstrated significant allele dose-response. We also found two SNPs that altered the 25(OH)D-relapse association with significant allele dose-response ($p_{interaction}$ <0.05), and five SNPs that modified levels of 25(OH)D (p_{trend} <0.05), with significant cumulative genotype 'risk' effect, and three demonstrated significant allele dose-response.

Given the nature of prospective cohort design, the genetic associations with clinical course can best be evaluated in longitudinal studies. However, the associations with genome-wide significance or significance after multiple testing requires a very large sample size and/or follow-up, which is prohibitively difficult for studies of MS clinical course. Therefore, using other methodologies to overcome potential type-1 error is necessary. In these studies, we conducted allele dose-responses and cumulative genotype effects analysis, and our findings support the nominal SNP associations and provide evidence for an association between known MS risk-associated SNPs and relapse. Therefore we believe that in studies of this type where sample size effects significantly

hinder the ability to detect statistically significant genetic associations, the utilisation of this ancillary supporting evidence to provide additional support is a worthwhile analysis method. These associations can then be studied in other cohorts as *a priori* hypotheses without the need for multiple testing limitations.

Using the same prospective MS cohort design, we also studied whether interactions between serum vitamin D and genes involved in the vitamin D pathway influence MS clinical course. We found two novel intronic SNPs in *PRKCZ* and *PRKCH* that significantly modify the association between serum 25(OH)D and hazard of relapse even after adjustment for multiple testing ($p_{interaction}$=0.001, P_{adj} =0.021 for both). We also identified one intronic SNP in *CYP2R1* and one novel intronic SNP in *PRKCB* that were significantly associated with 25(OH)D levels ($p_{interaction}$=0.001, P_{adj}=0.021 for both) after adjustment for multiple testing, and further with a cumulative effect of multiple 'risk' genotypes (p_{trend}= 7.12×10^{-6}). Among the identified SNPs, three locate in the protein kinase C (PKC) family genes, which have been shown to play critical roles in motor activity, development of the nervous system, and in learning and memory [26]. It has been demonstrated that PKC family genes are associated with other neurological disorders including Alzheimer's disease, status epilepticus and cerebellar ataxia [26]. The underlying mechanisms of PKC are not well understood. Research showed PKC plays a vital role in the regulation of signal transduction, cell proliferation and differentiation through positive and negative regulation of the cell cycle [27]. The eleven-member PKC family has been demonstrated to affect T-cell activation, such that *PRKCA* plays a crucial role in determining the magnitude of the T-cell proliferative response upon T-cell activation [28], and *PRKCB* has been suggested to be a key element in proper T cell migration [29]. Additionally both *PKCZ* and *PRKCE* have been implicated as having a

role in *IL-2* signaling in a murine T-cell line [30], and *PRKCQ* was involved in controlling the proliferation and differentiation of T-cells [31]. Interestingly, MS is believed to be the result of a misdirected immune response by autoreactive T-cells against as yet undetermined CNS antigens [32]. Therefore, modulation of these genes may be an important potential avenue of investigation in the treatment of MS and other immunological disorders.

Recently, PKCs have been suggested as potential therapeutic targets for treating diabetic complications, and oncological, inflammatory, immunological and dermatological disorders [27]. For example, biological agents that affect the activity of *PRKCQ* are being developed and tested for their potential as novel therapy for several T-cell-mediated disease conditions such as MS, rheumatoid arthritis, asthma, and inflammatory bowel disease, and to assist in organ transplantation [31], via targeting effector T-cells (Teffs) and regulatory T-cells (Tregs) to reduce and conversely increase their in vivo survival and function [33]. Interestingly, several PKC family genes have been demonstrated to be associated with infectious and autoimmune disease, such that *PRKCA* was associated with MS risk in UK, Finish and Canadian populations [34,35], and *PRKCB* and *PRKCH* had been associated with the risk of similar autoimmune diseases including systemic lupus erythematosus (SLE) [36] and rheumatoid arthritis [37,38]. Although the exact underlying mechanisms of these PKC isoforms are not well understood, these findings indicated that they might be potential therapies to MS and other autoimmune disease, as with *PRKCQ*. However, different PKC isoforms have high sequence identity but they are involved in different diseases [39]. Each PKC isoform performs unique function due to its isoform-specific patterns of subcellular compartmentalization, protein-protein interactions and post-translational modifications [39]. In addition, the clinical trial candidates of PKCs

mainly target the catalytic domain, which is highly conserved throughout the PKC family and which make it difficult to target a particular isoform selectively [27]. Thus, identification of the crystal structures and thorough analysis of available structures and information on the PKC ligands will be helpful in drug design and development processes [39]. At present, few crystal structures of PKC isoforms are known [39] and further research into the crystal structure of PKC family genes is required. In the current study, we found that the identified PKC family genes do not have strictly independent roles but work in conjunction with each other, suggesting cooperative mechanisms among PKC family genes. This potential for cooperative effects requires further investigation.

Overall, our findings in Chapter 4 and Chapter 5 have indicated for the first time that gene-vitamin D interactions may be an important mechanism in MS clinical course, and that the PKC family genes may play a role in the pathogenesis of MS relapse through modulating the association between 25(OH)D and relapse. First, replication in other independent dataset is yet required, ideally via high-quality prospective cohort designs. Secondly, understanding the mechanism of action of PKCs in disease conditions will be helpful in finding new and potential PKC modulators. The underlying mechanisms of which PKC family genes and how they interact with serum vitamin D, and how they work in conjunction require further investigation. Third, the crystal structures of the PKC family genes need to be dissected, and consequently develop and test PKC as therapeutic agents of MS or other immunological disorders in the long run.

7.3 Gene-IFN-β interactions and serum vitamin D

Chapter 6 described how to identify the interactions between IFN-β and those genes involved in the vitamin D pathway influence serum vitamin D levels. We found that two

novel intronic SNPs in Wilms' Tumor 1 (*WT1*) gene significantly modified the previously demonstrated association between IFN-β and 25(OH)D in people with MS and RRMS, and showed evidence of modulating the sun-IFN-β-25(OH)D association and the 25(OH)D-IFN-β-relapse association. These findings indicate that *WT1* may play an important role in altering the therapeutic effects of IFN-β in MS. Subject to validation, these results potentially implicate *WT1* as a pharmacogenomic target in MS, though clearly more work is needed to delineate the mechanism at the cellular level.

WT1 gene encodes a zinc finger DNA-binding protein that acts as a transcriptional activator or repressor, and plays an essential role in the normal development of the urogenital system. It is well known that *WT1* gene is associated with Wilms' tumor [40-42] onset and a variety of urinary system-related disease [43-47]. *WT1* has also been associated with Alzheimer's disease [48], as elevations of *WT1* correlate with increased levels of apoptosis, which has been suggested to play a role in neurodegeneration in Alzheimer's disease.

Little is known about *WT1* on MS and on the therapeutic effect of IFN-β, however. It is known that IFN-β is one of the first-line treatments for RRMS. The therapeutic mechanism of IFN-β is still not fully understood, but it is now thought that it binds to receptors on immune cells and initiates a complex transcriptional response, and it inhibits antigen presentation and T-cell proliferation with reduced cytokine and matrix metalloproteinase expression [49]. Notably, one study demonstrated that co-administration of *WT1* peptide and IFN-β enhanced tumor immunity, mainly through the induction of *WT1*-specific cytotoxic T-lymphocytes (CTLs), enhancement of natural killer cell activity, and promotion of MHC class I expression on tumor cells [50]. This study

suggested that *WT1* was involved in immune activity, mediated by CTLs and consequently may also alter the therapeutic effect of IFN-β. *WT1* has also been demonstrated to modulate the response to active form of vitamin D (1,25-dihydroxyvitamin D3) by induction of VDR [51]. In addition, the human VDR gene has been confirmed as a downstream target of *WT1*, with three *WT1*-responsive sites identified in the *VDR* promoter [52]. Interestingly, *WT1* is involved in cell proliferation as one cofactor with the *PDGFA* gene and also binds to VDR in the downstream section of the vitamin D pathway (**Figure 6.1** in this thesis). These findings suggest *WT1* expression may be mediated by VDR in downstream signaling pathways, consequently influencing the level of serum 25(OH)D. However, the mechanism by which serum 25(OH)D levels are impacted by transcriptional activation of VDR by *WT1* is unclear. Taken together these results are enticing; however more investigation into the mechanisms of how *WT1* is involved in immune activation and altering IFN-β's effect in MS is still needed.

In summary, we demonstrated first time in Chapter 6 that the *WT1* gene modified the relationship between IFN-β and 25(OH)D in people with MS and RRMS. These findings firstly need to be ereplicate in other independent studies, ideally via prospective cohort design. Furthermore, the modulating effects of *WT1* may be mediated by immune system or by VDR in signaling pathways. The underlying mechanisms of how *WT1* involved in immune activation with IFN-β and how WT1 involved in transcriptional activation on VDR need further investigation. In addition, we have demonstrated previously that IFN-β treatment was associated with higher 25(OH)D levels among persons with MS, this apparently mediated by a stronger association between personal sun exposure and subsequently measured 25(OH)D [53]. The observed modulating effect of *WT1* on the relationship between IFN-β and 25(OH)D maybe also mediated by sun exposure, since

evidence provided from the study that *WT1* modified the relationship between sun exposure and IFN-β in predicting 25(OH)D. Therefore, more research into *WT1* on the relationship among sun exposure, IFN-β and 25(OH)D is needed.

7.4 Final conclusions of PhD

Briefly, my PhD research focuses on how to dissect the genetic architecture of MS in two ways. One way is to discover genetic variants associated with MS, including detecting rare disease-causing variants that confer susceptibility on MS. Another way is to discover gene-environment interactions associated with MS, including detecting gene-vitamin D interactions and gene- IFN-β interactions in patients with MS.

By utilising IBD mapping - BEAGLE fastIBD with a big GWAS dataset, we identified a genome-wide significant linkage signal on chromosome 19. The genes in the linkage region may be involved in early developmental regulation of the CNS. Our findings suggested BEAGLE fastIBD is a good way to detect rare variants in large unrelated population.

By utilising family-based whole genome sequencing analysis, we identified nine candidate genes may be disease-causing variants of MS. Where two novel coding variants located in *RDBP* and *PKHD1* suggest to be promising potential disease-causing variants to MS, which will be validated in the further research. Although analysis is ongoing, the initial findings indicated that family-based whole genome sequencing analysis is a promising alternative to detect disease-causing variants.

By utilising a prospective cohort design in the Southern Tasmanian Multiple Sclerosis Longitudinal Study, we demonstrated that gene-vitamin D interactions influence MS clinical course, such that a number of known MS risk-associated genetic variants and a number of novel *PKC* family genes significantly modified the relationship between 25(OH)D and relapse. In addition, we also demonstrated that a number of known MS risk-associated genetic variants and *PKC* family genes were also associated with MS relapse and 25(OH)D levels. These findings indicated for the first time that *PKC* family genes may play an important role in the pathogenesis of MS relapse and their effects on relapse may be modulated by vitamin D levels. By using the same longitudinal study, we also demonstrated for the first time that gene- IFN-β interactions influence serum vitamin D levels, such that *WT1* gene significantly modified the relationship between IFN-β and 25(OH)D. These findings may suggest a genetic basis for the significant pharmacological differences seen between cases treated with IFN-β. If supported by replication studies, these findings may provide the basis for determining potential responders or non-responders to IFN-β therapy, thereby allowing clinicians to select the appropriate treatment, that is, those who are more likely to respond to therapy and not exposing those who may be better suited to other medications.

7.5 Future research

Future research should initially focus on replications of our findings. In the case of IBD mapping to detect rare variants, this will preferably use family-base sequencing data or population-based next-generation sequencing data or ideally whole-genome sequencing data to replicate it. We are considering to use the pedigree-based sequencing data comprised a large Mexican family from Biomedical Research Institute, Texas, USA in the future. As for assessing disease-causing variants of MS by using family-based whole

genome sequencing analysis, further research will involve testing the two most interesting candidate variants in as many MS cases and controls as can be recruited. This work has already commenced and I will provide significant input into future work based on my PhD thesis. Furthermore, raw whole-genome sequencing data will be reanalysed using pedigree-based pipeline to detect missing vital rare variants in the coming days.

The results of gene-vitamin D interactions influencing MS clinical course need to be replicated in other high-quality longitudinal MS studies, preferably with lager sample sizes and longer follow-up. These types of studies are very hard to conduct, expensive and highly demanding on the participants. I am unaware of any similar longitudinal study to the MSL Study anywhere in the world, making these findings difficult to replicate by such means. On the other hand, the mechanisms of how *PKC* family genes interact with serum vitamin D require further investigation, particularly from the cellular level. Since researchers have identified *PKC* family genes products as potential therapeutic targets for treating inflammatory and immunological disorders, the crystal structures of *PKC* family genes need to be dissected further, and the identification of novel PKC inhibitors have been conducted by several pharmaceutical companies and institutes [54-59]. These proteins or agonists/antagonists studied in vitro need to develop and test PKC as a therapeutic agent in MS or other immunological disorders, and evenly apply in the clinical realm.

Similarly, the results of gene-IFN-β interactions influencing serum vitamin D levels need to be replicated in other independent datasets, preferably via longitudinal studies, with large sample size and long follow-up. In addition, the mechanisms of how *WT1* alter the therapeutic effect of IFN-β need further investigation, including the mechanisms of how

WT1 is involved in immune activation and how *WT1* is involved in transcriptional activation of VDR in the vitamin D signaling pathway in patients of MS, preferably via cellular level investigations.

If validated in other studies the next step would be to study these genes using functional genomics either in silica or in vitro. These findings open up a significant new frontier for MS research and demonstrate that the hunt for genetic factors that influence MS clinical course can be conducted in well thought out studies of MS clinical course without requiring vast numbers of cases. Utilising the methods as we have that were described by Bradford-Hill (the Hill criteria [60]) we have been able to ascribe causality within the framework of our research studies despite the limitations of sample size and multiple testing burden. I believe that these findings, if replicated, can be rapidly introduced into the basic science sphere of research to elucidate the cellular mechanism by which these gene-environment interactions affect MS clinical course.

7.6 References

1. Jin Y, Qu S, Tesikova M, Wang L, Kristian A, et al. (2013) Molecular circuit involving KLK4 integrates androgen and mTOR signaling in prostate cancer. Proc Natl Acad Sci U S A 110: E2572-2581.
2. Tan Z, Zhang S, Li M, Wu X, Weng H, et al. (2013) Regulation of cell proliferation and migration in gallbladder cancer by zinc finger X-chromosomal protein. Gene.
3. Awata T, Yamashita H, Kurihara S, Morita-Ohkubo T, Miyashita Y, et al. (2013) A low-frequency GLIS3 variant associated with resistance to Japanese type 1 diabetes. Biochem Biophys Res Commun.
4. Suliman BA, Xu D, Williams BR (2012) The promyelocytic leukemia zinc finger protein: two decades of molecular oncology. Front Oncol 2: 74.
5. Zhong J, Chen S, Xue M, Du Q, Cai J, et al. (2012) ZIC1 modulates cell-cycle distributions and cell migration through regulation of sonic hedgehog, PI(3)K and MAPK signaling pathways in gastric cancer. BMC Cancer 12: 290.
6. Cheng Y, Liang P, Geng H, Wang Z, Li L, et al. (2012) A novel 19q13 nucleolar zinc finger protein suppresses tumor cell growth through inhibiting ribosome

biogenesis and inducing apoptosis but is frequently silenced in multiple carcinomas. Mol Cancer Res 10: 925-936.
7. Yue X, Zhang Z, Liang X, Gao L, Zhang X, et al. (2012) Zinc fingers and homeoboxes 2 inhibits hepatocellular carcinoma cell proliferation and represses expression of Cyclins A and E. Gastroenterology 142: 1559-1570 e1552.
8. Richter S, Morrison S, Connor T, Su J, Print CG, et al. (2013) Zinc Finger Nuclease Mediated Knockout of ADP-Dependent Glucokinase in Cancer Cell Lines: Effects on Cell Survival and Mitochondrial Oxidative Metabolism. PLoS One 8: e65267.
9. Day T, Bianco-Miotto T (2013) Common gene pathways and families altered by DNA methylation in breast and prostate cancer. Endocr Relat Cancer.
10. Huang RL, Chang CC, Su PH, Chen YC, Liao YP, et al. (2012) Methylomic analysis identifies frequent DNA methylation of zinc finger protein 582 (ZNF582) in cervical neoplasms. PLoS One 7: e41060.
11. Wu S, Lao XY, Sun TT, Ren LL, Kong X, et al. (2013) Knockdown of ZFX inhibits gastric cancer cell growth in vitro and in vivo via downregulating the ERK-MAPK pathway. Cancer Lett.
12. Yu J, Liang QY, Wang J, Cheng Y, Wang S, et al. (2013) Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer. Oncogene 32: 307-317.
13. Okugawa Y, Toiyama Y, Tanaka K, Matsusita K, Fujikawa H, et al. (2012) Clinical significance of Zinc finger E-box Binding homeobox 1 (ZEB1) in human gastric cancer. J Surg Oncol 106: 280-285.
14. Izumi H, Wakasugi T, Shimajiri S, Tanimoto A, Sasaguri Y, et al. (2010) Role of ZNF143 in tumor growth through transcriptional regulation of DNA replication and cell-cycle-associated genes. Cancer Sci 101: 2538-2545.
15. Gan L, Chen S, Zhong J, Wang X, Lam EK, et al. (2011) ZIC1 is downregulated through promoter hypermethylation, and functions as a tumor suppressor gene in colorectal cancer. PLoS One 6: e16916.
16. Yu T, Chen X, Zhang W, Colon D, Shi J, et al. (2012) Regulation of the potential marker for intestinal cells, Bmi1, by beta-catenin and the zinc finger protein KLF4: implications for colon cancer. J Biol Chem 287: 3760-3768.
17. Lleras RA, Adrien LR, Smith RV, Brown B, Jivraj N, et al. (2011) Hypermethylation of a cluster of Kruppel-type zinc finger protein genes on chromosome 19q13 in oropharyngeal squamous cell carcinoma. Am J Pathol 178: 1965-1974.
18. Wang Q, Tan YX, Ren YB, Dong LW, Xie ZF, et al. (2011) Zinc finger protein ZBTB20 expression is increased in hepatocellular carcinoma and associated with poor prognosis. BMC Cancer 11: 271.
19. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. Am J Hum Genet 88: 173-182.
20. Su SY, Kasberger J, Baranzini S, Byerley W, Liao W, et al. (2012) Detection of identity by descent using next-generation whole genome sequencing data. BMC Bioinformatics 13: 121.
21. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.
22. Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, et al. (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27: 829-836.
23. Su SY, Kasberger J, Baranzini S, Byerley W, Liao W, et al. (2012) Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. BMC Bioinformatics 13: 121.

24. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851-862.
25. Schymick JC, Scholz SW, Fung HC, Britton A, Arepalli S, et al. (2007) Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol 6: 322-328.
26. Kodavanti PR (2011) Cell signaling and neurotoxicity: protein kinase C in vitro and in vivo. Methods Mol Biol 758: 307-319.
27. Sobhia ME, Grewal BK, Ml SP, Patel J, Kaur A, et al. (2013) Protein kinase C inhibitors: a patent review (2008 - 2009). Expert Opin Ther Pat.
28. Baier G (2003) The PKC gene module- molecular biosystematics to resolve its T cell functions. Immunol Rev 192: 64-79.
29. Volkov Y, Long A, McGrath S, Ni Eidhin D, Kelleher D (2001) Crucial importance of PKC- (I) in LFA-1-mediated locomotion of acti- vated T cells. Nat Immunol 2: 508-514.
30. Gomez J, Pitton C, Garcia A, Martinez de Aragon A, Silva A, et al. (1995) The zeta isoform of protein kinase C controls interleukin-2-mediated proliferation in a murine T cell line: evidence for an additional role of protein kinase C epsilon and beta. Exp Cell Res 218: 105-113.
31. Chand S, Mehta N, Bahia MS, Dixit A, Silakari O (2012) Protein kinase C-theta inhibitors: a novel therapy for inflammatory disorders. Curr Pharm Des 18: 4725-4746.
32. Noseworthy JH, Lucchinetti C, Rodriguez M, Weinshenker BG (2000) Multiple sclerosis. N Engl J Med 343: 938–952.
33. Zanin-Zhorov A, Dustin ML, Blazar BR (2011) PKC-theta function at the immunological synapse: prospects for therapeutic targeting. Trends Immunol 32: 358-363.
34. Barton A, Woolmore JA, Ward D, Eyre S, Hinks A, et al. (2004) Association of protein kinase C alpha (PRKCA) gene with multiple sclerosis in a UK population. Brain 127: 1717-1722.
35. Saarela J, Kallio SP, Chen D, Montpetit A, Jokiaho A, et al. (2006) PRKCA and multiple sclerosis: association in two independent populations. PLoS Genet 2: e42.
36. Sheng YJ, Gao JP, Li J, Han JW, Xu Q, et al. (2011) Follow-up study identifies two novel susceptibility loci PRKCB and 8p11.21 for systemic lupus erythematosus. Rheumatology (Oxford) 50: 682-688.
37. Takata Y, Hamada D, Miyatake K, Nakano S, Shinomiya F, et al. (2007) Genetic association between the PRKCH gene encoding protein kinase Ceta isozyme and rheumatoid arthritis in the Japanese population. Arthritis Rheum 56: 30-42.
38. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet 44: 511-516.
39. Sobhia ME, Grewal BK, Paul ML, Patel J, Kaur A, et al. (2013) Protein kinase C inhibitors: a patent review (2010 - present). Expert Opin Ther Pat.
40. Royer-Pokora B, Beier M, Henzler M, Alam R, Schumacher V, et al. (2004) Twenty-four new cases of WT1 germline mutations and review of the literature: genotype/phenotype correlations for Wilms tumor development. Am J Med Genet A 127A: 249-257.
41. Sakamoto J, Takata A, Fukuzawa R, Kikuchi H, Sugiyama M, et al. (2001) A novel WT1 gene mutation associated with wilms' tumor and congenital male genitourinary malformation. Pediatr Res 50: 337-344.

42. Gessler M, Konig A, Moore J, Qualman S, Arden K, et al. (1993) Homozygous inactivation of WT1 in a Wilms' tumor associated with the WAGR syndrome. Genes Chromosomes Cancer 7: 131-136.
43. Hinkes BG, Mucha B, Vlangos CN, Gbadegesin R, Liu J, et al. (2007) Nephrotic syndrome in the first year of life: two thirds of cases are caused by mutations in 4 genes (NPHS1, NPHS2, WT1, and LAMB2). Pediatrics 119: e907-919.
44. Pelletier J, Bruening W, Kashtan CE, Mauer SM, Manivel JC, et al. (1991) Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys-Drash syndrome. Cell 67: 437-447.
45. Ghahremani M, Chan CB, Bistritzer T, Aladjem MM, Tieder M, et al. (1996) A novel mutation H373Y in the Wilms' tumor suppressor gene, WT1, associated with Denys-Drash syndrome. Hum Hered 46: 336-338.
46. Orloff MS, Iyengar SK, Winkler CA, Goddard KA, Dart RA, et al. (2005) Variants in the Wilms' tumor gene are associated with focal segmental glomerulosclerosis in the African American population. Physiol Genomics 21: 212-221.
47. Wang Y, Li Q, Xu J, Liu Q, Wang W, et al. (2004) Mutation analysis of five candidate genes in Chinese patients with hypospadias. Eur J Hum Genet 12: 706-712.
48. Lovell MA, Xie C, Xiong S, Markesbery WR (2003) Wilms' tumor suppressor (WT1) is a mediator of neuronal degeneration associated with the pathogenesis of Alzheimer's disease. Brain Res 983: 84-96.
49. Markowitz CE (2007) Interferon-beta: mechanism of action and dosing issues. Neurology 68: S8-11.
50. Nakajima H, Oka Y, Tsuboi A, Tatsumi N, Yamamoto Y, et al. (2012) Enhanced tumor immunity of WT1 peptide vaccination by interferon-beta administration. Vaccine 30: 722-729.
51. Maurer U, Jehan F, Englert C, Hubinger G, Weidmann E, et al. (2001) The Wilms' tumor gene product (WT1) modulates the response to 1,25-dihydroxyvitamin D3 by induction of the vitamin D receptor. J Biol Chem 276: 3727-3732.
52. Lee TH, Pelletier J (2001) Functional characterization of WT1 binding sites within the human vitamin D receptor gene promoter. Physiol Genomics 7: 187-200.
53. Stewart N, Simpson SJ, van der Mei I, Ponsonby AL, Blizzard L, et al. (2012) Interferon-β and serum 25-hydroxyvitamin D interact to modulate relapse risk in MS. Neurology 79: 254-260.
54. Boehringer Ingelheim International GmbH. (2008) Pyridine derivatives useful as inhibitors of PKC-theta. US7468382B2.
55. Novartis AG (2008) Indolylmaleimide derivatives as PKC inhibitors. US20080242675A1.
56. Pfizer I (2008) 3-amino-pyrrolo [3, 4-C] pyrazole- 5 (1H, 4H, 6H) carbaldehyde derivatives as PKC inhibitors. WO2008096260.
57. Boehringer Ingelheim Pharma Gmbh & Co. Kg. (2009) Pyrimidine derivatives useful as inhibitors of PKC-theta
. US7,550,473B2.
58. Boehringer Ingelheim Pharmaceuticals I (2009) Pyrimidine derivatives useful as inhibitors of PKC-theta. US7601714B2.
59. Janssen Pharmaceutica N.V. (BE). (2009) Substituted Pyrroline Kinase Inhibitors. US7488826B2.
60. Hill AB (1965) The Environment and Disease: Association or Causation? Proc R Soc Med 58: 295-300.

The following article has been removed for copyright or proprietary reasons

Appendix 7A: other publications during PhD

O'Gorman C, **Lin R,** Stankovich J, Broadley SA. Modelling genetic susceptibility to multiple sclerosis with family data. *Neuroepidemiology*. 2013; 40(1):1-12.